there is no need to rewrite them in independent form, because, for the reasons noted *infra*, the claims from which they depend are patentable.

The objection to claim 18 is respectfully traversed in view of the cancellation of this claim.

The rejection of claims 9, 11, 18-20, 23, 27, 29, 34, 38, 40, 45, 49, 51, 56, and 63 under 35 U.S.C. § 112, first paragraph, for lack of enablement is respectfully traversed in view of the above amendments.

The rejection of claims 69 and 70 under 35 U.S.C. § 112, first paragraph, for lack of enablement is obviated in view of the cancellation of these claims.

The rejection of claims 5-7, 9, 12, 13, 25, 30, 31, 36, 41, 42, 47, 52, 53, 59, 64, and 65 under 35 U.S.C. § 112, first paragraph, for lack of written description is respectfully traversed in view of the above amendments. Claims 9, 25, 30, 31, 36, 41, 42, 47, 52, and 53 have been canceled. Further, applicant has limited independent claims 5 and 59 to "[a]n isolated DNA molecule encoding a protein subunit of polymerase III holoenzyme from a eubacterial prokaryote" and independent claims 14 and 54 "[a]n isolated protein subunit of polymerase III holoenzyme from a eubacterial prokaryote". The disclosure of the DNA molecule encoding the δ' and δ subunits for *E. coli* in the present application is a representative DNA species of the protein subunit of polymerase III holoenzyme from a eubacterial prokaryote and, therefore, applicant has provided written description for the independent claims and any claims which depend therefrom.

It is well known to one skilled in the art that proteins homologous to the δ' subunit of the *E. coli* polymerase III holoenzyme are contained in organisms other than *E. coli*, as shown in the Declaration of Michael O'Donnell under 37 CFR § 1.132 submitted in parent U.S. Patent Application Serial No. 08/279,058 on December 17, 1996 ("Supp. O'Donnell Declaration") and the Supplemental Declaration of Michael O'Donnell under 37 CFR § 1.132 ("Supp. O'Donnell Declaration") (submitted herewith).

Those skilled in the art recognize the δ' subunit from *E. coli* has sequence homology to accessory protein complexes of various other organisms (O'Donnell Declaration ¶ 13). For example, in O'Donnell et al., "Homology in Accessory Proteins of Replicative Polymerases - *E. coli* to Humans," Nucleic Acids Research 21(1):1-3 (1993), a comparison of amino acid sequences shows the homology between proteins of replicative polymerases of *E. coli*, humans, and phage T4 (Id.). In Carter et al., "Identification, Isolation, and Characterization of the Structural Gene Encoding the δ' Subunit of *Escherichia coli* DNA

Polymerase III Holoenzyme," J. of Bacteriology, 175(12):3812-22 (1993), Figure 5 diagrams the homology of the δ' amino acid sequence to other replication proteins (Id.). Comparison of the δ' amino acid sequence revealed similarity to the A1(replication factor C) complex of HeLa cells and to the gene 44 protein (gp44) of bacteriophage T4 (Id.). In addition, amino acid sequence similarity was found to the gene product of *B. subtilis* (Id.). Further, the structural homology of the δ' subunit to other replication proteins has been proven to be true (Id.). For example, the genome project of *Haemophilus influenzae* showed homologues to all 10 subunits of *E. coli* DNA polymerase III holoenzyme, including δ, δ', χ, Ψ, and θ (Id.). Currently, the GenBank now also shows homologues to the δ' subunit of *E. coli* from a large variety of organisms, including the following prokaryotes: *Escherichia coli, Haemophilus influenzae, Micrococcus luteus, Pseudomonas aeruginosa, Bacillus subtilis,* and *Caulobacter crescentus* (Id.).

As to the δ and δ' subunits of polymerase III holoenzyme, it is well known to one skilled in the art that proteins in other organisms have functional and structural homology to the subunits of *E. coli* (O'Donnell Declaration ¶¶ 10-16).

Various genome projects for many different organisms have resulted in the gene sequences for various bacteria being publicly available on various web sites (Supp. O'Donnell Decl. ¶5). As described more fully below, the amino acid sequences for the δ and δ' subunits for *E. coli*, disclosed in the present application, were used, by myself and others, in a BLAST search program (Altschul, et al., "Basic Local Alignment Search Tool," J. Mol. Bio. 215:403-10 (1990)) to identify the presence of genes encoding these proteins in other eubacterial prokaryotes (Id.). As explained in the textbook Molecular Biology of the Gene (attached to the Supp. O'Donnell Decl. at Appendix A), eubacterial (i.e. true bacteria) prokaryotes are a distinct kingdom separate from eukaryotes and archaebacteria and include: Aquificales (included *Aquifex aeolicus*), *Chlamydiales*, *Coprothermobacter*, *Cyanobacteria*, Green Sulfur bacteria (includes *Porphyromonal gingivalis* and *Chlorobium tepidum*), *Fibrobacter* group, *Firmicutes* (Gram positives including *Mycobacterium, Clostridium acetobutylicum*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Staphylococcus aureus*, *Bacillus subtilis*), *Flexistipes* group, *Fusobacteria*, Green non-sulfur bacteria, *Holophaga* group, *Nitrospira* group, *Planctomycetales*, *Proteobacteria* (includes the alpha subdivision (e.g. *Caulobacter crescentus*), the beta group (e.g. *Bordetella pertussis* and *Neisseria meningitidis*), the delta/epsilon subdivisions (e.g. *Campylobacter jejuni* and *Helicobacter pylori*), and the gamma subdivision (e.g. the *Enterobacteriaceae* that includes *Haemophilus*

influenzai, *Yersinia pestis*, *Vibrio cholerai*, *Escherichia coli*, *Pasturella multocida*, *Pseudomonas aeruginosa*, *Salmonells typhi*, *Shewanella putrefaciens*), *Spirochaetales* (includes *Borrelia burgdorgeri*, *Treponema palladum*), *Synergistes* group, *Thermodesulfobacterium* group, *Thermotogales* (included *Thermotoga maritima*), *Thermus/Deinococcus* group (included *Thermus thermophilis* and *Deinococcus radiodurans*), and a variety of as yet unclassified bacteria. The results of these analyses are set forth below (Id.).

The sequence analysis of *Haemophilus influenze* is found at http://www.tigr.org/tdb/mdb/hidb/hidb.html (Supp. O'Donnell Decl. ¶6). A copy of that web site listing is attached to the Supp. O'Donnell Decl. at Appendix B with the δ subunit encoding gene being identified as HI0923 and the δ' subunit encoding DNA molecule being identified as HI0455 (Id.). This listing shows that the δ subunit encoding DNA molecule of *Haemophilus influenze* is 62.0% similar to the δ subunit encoding DNA molecule of *E. coli* (Id.). Likewise, the δ' subunit of *Haemophilus influenze* is shown to be 57.4% similar to the δ' subunit encoding DNA molecule of *E. coli* (Id.).

The genome of *Niceria gonorrhoeae* is found at the web site http://www.genome.on.edu (Supp. O'Donnell Decl. ¶7). Search for the δ subunit amino acid sequence yields a contig. with a very high probability of 1.2×10^{-25}, contig. 188, while the δ' amino acid sequence yields a contig. of high probability of 1.2×10^{-14} #200 (Id.). See Appendix C attached to the Supp. O'Donnell Decl.

The genome for *Shewanella putrefaciens* is found on the TIGR BLAST server (Supp. O'Donnell Decl. ¶8). A search for the δ subunit produced the high score of 1.1×10^{-54} for contig. gsp 230, while the search for δ' subunit produced the high score of 6.4×10^{-27} for contig. gsp 271 (Id.). See Appendix D attached to the Supp. O'Donnell Decl.

The genome for *Vibrio cholerae* is found at http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=v.cholerae (Supp. O'Donnell Decl. ¶9). A search for the δ subunit produced the high score of 6.9×10^{-82} for contig. asm 937, while the search for δ' subunit produced the high score of 8.1×10^{-37} for contig. asm 894 (Id.). See Appendix E attached to the Supp. O'Donnell Decl.

The genomes for *Pseudomonas aeruginosa* (see Appendix F attached to the Supp. O'Donnell Decl.), *Salmonella typhi* (see Appendix G attached to the Supp. O'Donnell Decl.), and *Yersinia pestis* (see Appendix H attached to the Supp. O'Donnell Decl.) are found at http://www.ncbi.nlm.nih.gov/Blast/unfinished genomes (Supp. O'Donnell Decl. ¶10). For

these, the amino acid sequences of *E. coli* δ and δ' were used in BLAST searches (Id.). The high scores, given below, are all sufficiently significant to call the identified gene the one that performs the homologous function in *E. coli* (Id.):

Pseudomonas aeruginosa
δ - 7 x 10^{-34} contig. 52
δ' - 9 x 10^{-27} contig. 50

Salmonella typhi
δ - 1 x 10^{-161} contig. 1564
δ' - 8 x 10^{-10} contig. 870

Yersinia pestis
δ - 1 x 10^{-127} contig. 803
δ' - 9 x 10^{-98} contig. 51

Thus, for Gram negative bacteria such as *Haemophilus influenze*, *Niceria gonorrhoeae*, *Shewanella putrefaciens*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Salmonella typhi*, and *Yersinia pestis*, there is a high level of homology between the δ and δ' subunits of those bacteria and the δ and δ' subunits of *E. coli* (Supp. O'Donnell Decl. ¶11).

For other eubacteria, there is significant homology between their δ' subunit and that of *E. coli* (Supp. O'Donnell Decl. ¶12). In all eubacteria, the δ subunit can be identified starting with the *E. coli* δ subunit as comparison, but, since it is not as conserved as the δ' subunit, one must "walk" from one organism to another, as discussed in ¶ 23 below (Id.).

In Himmelreich et al., "Complete Sequence Analysis of the Genome of the Bacterium *Mycoplasma pneumoniae*," Nucleic Acids Research 24(22):4420-4449 (1996), the δ' subunit of *Mycoplasma pneumoniae* is identified as being homologous to the δ' subunit of *E. coli* in Table 1 on page 4426 (Supp. O'Donnell Decl. ¶13). See Appendix I attached to the Supp. O'Donnell Decl.

In Kunst et al., "The Complete Genome Sequence of the Gram-positive Bacterium *Bacillus subtilis*," Nature 390:249-256 (1997), the δ' subunit of *Bacillus subtilis* is identified as being homologous to the δ' subunit of *E. coli* in the table on page 248 (Supp. O'Donnell Decl. ¶14). See Appendix J attached to the Supp. O'Donnell Decl.

The genome for *Streptococcus pyogenes* is found in the University of Oklahoma server (i.e. http://www.ncbi.nlm.nih.gov.BLAST/tigrbl.html) (Supp. O'Donnell

Decl. ¶15). δ' produced the high score of 3.3 x 10^{-10} for contig. 218 (Id.). See Appendix K attached to the Supp. O'Donnell Decl.

The genome for *Enterococcus faecalis* is found on the TIGR BLAST search server (Supp. O'Donnell Decl. ¶16). δ' produced the high score of 9.6 x 10^{-16} for contig. 6277 (Id.). See Appendix L attached to the Supp. O'Donnell Decl.

The genome for *Streptococcus pneumoniae* is found on the TIGR BLAST search server (Supp. O'Donnell Decl. ¶17). δ' produced the high score of 2.4 x 10^{-12} for contig. sp 68 (Id.). See Appendix M attached to the Supp. O'Donnell Decl.

The genome for *Aquifex aeolicus* is found in Deckert et al., "The Complete Genome of the Hyperthermophilic bacterium *Aquifex aeolicus*," Nature 392:353-358 (1998) and at http://www.ncbi.nlm.nih.gov/Blast/unfinished genomes (Supp. O'Donnell Decl. ¶18). δ' produced the high score of 8 x 10^{-13} (position 1303996-1304394) (Id.). See Appendix N attached to the Supp. O'Donnell Decl.

The genome for *Thermatoga maritima* is found in the TIGR BLAST server page (Supp. O'Donnell Decl. ¶19). δ' yields a high score of 3.7 x 10^{-15} for contig. tm 26 (Id.). See Appendix O attached to the Supp. O'Donnell Decl.

In *Spirochaetes*, Tomb et al., "The Complete Genome Sequence of the Gastric Pathogen *Helicobacter pylori*," Nature 388:539-547 (1997) (see Appendix P attached to the Supp. O'Donnell Decl.) and Fraser et al., "Genomic Sequence of a Lyme Disease Spirochaete, *Borrelia burgdorferi*," Nature 390:580-586 (1997) (see Appendix Q attached to the Supp. O'Donnell Decl.), *Helicobacter pylori* and *Borrelia burgdorferi* are identified to have δ' subunits (Supp. O'Donnell Decl. ¶20). For *Helicobacter pylori*, δ' is listed in the table as HP1231 (Id.). For *Borrelia burgdorferi*, using the NCBI genome search page (Ncbi.nlm.nih.gov/Blast/unfinished genomes), δ' gives the high score of 8 x 10^{-7} (Id.). See Appendix R attached to the Supp. O'Donnell Decl.

In Andersson et al., "The Genome Sequence of *Rickettsia prowazekii* and the Origin of Mitochondria," Nature 396:133-140 (1998), *Rickettsia prowazekii* is identified to have a δ' subunit, identified as RP172 (Supp. O'Donnell Decl. ¶21). See Appendix S attached to the Supp. O'Donnell Decl.

A large compilation of genome sequences is at the web site http://www.ncbi.nlm.gov/Blast/unfinished genome.html (Supp. O'Donnell Decl. ¶22). The eubacterial genomes were searched using the δ' subunit of *E. coli* (Id.). All organisms in eubacteria scored very high with identity levels sufficient to identify the holB gene encoding

δ' conclusively (Id.). This is seen in Figure 1 showing a path of one-on-one comparative alignments each of which start with *E. coli* and the alignments (Id.) (Appendix T attached to the Supp. O'Donnell Decl.). In this figure, within the parentheses, is the percent identity and the ratio of the number of identities (i.e. the numerator) over the length of the amino acid sequence that was compared (i.e. the denominator) (Id.). The number outside of the parentheses is the score obtained in the Blast program (i.e. even a score of 1×10^{-9} is a sufficiently high score to identify the homologous gene) (Id.).

A similar search with the δ subunit of *E. coli* identified the holA gene of *Nisseria* and *Thiobacillus* as high matches, and holA of other enteric bacteria produced high scores as well (Supp. O'Donnell Decl. ¶23). Repetition of this procedure using *Neisseria* δ easily allows the identification of δ in *Aquifex aeolicus* (Id.). Use of *Aquifex aeolicus* δ identifies δ of *Enterococcus* (which identifies *Bacillus* δ, then *Streptococcus* δ, then *Synechocystis*, and the *Porphyromonas* δ) (Id.). Use of *Aquifex aeolicus* δ also identifies *Thermatoga* δ, which identifies Spirochaetes (*Borrelia*) δ subunit (Id.). Use of *Thiobacillus* δ identifies δ from *Helicobacter camylobacter* (Id.). There is a region at about 100 residues that is rather well conserved in δ across eubacteria and if this were used, the scores could be even higher yet (Id.). Figure 2 shows this "walking" procedure and shows the scores and percent identities obtained as a result of this procedure starting from the δ subunit of *E. coli* as well as alignments (Id.). This figure is substantially the same as Figure 1 but within the parentheses, after the percentage identity, there is another ratio and another percentage based on homologies (Id.). Figure 2 does not show scores for individual Gram negative bacteria of the *Enterobacteria* class (called enterics) as they are highly related to *E. coli* and the scores are very high (Id.).

Therefore, those of ordinary skill in the art, using the sequence information in the present application, would have been able to (and, in fact, did) identify and isolate the δ and δ' subunits of polymerase III holoenzyme (and their encoding genes) from eubacteria other than *E. coli* (See Supp. O'Donnell Declaration ¶ 24).

Further, the sequence of the eubacterial homologues to δ', and indeed the other δ' homologues, are sufficiently homologous to the δ' subunit of *E. coli* to provide for identifying and obtaining the corresponding δ' (holA) gene from these organisms using the gene encoding the δ' subunit of *E. coli* in the following ways: (1) use of the *E. coli* holA gene, or fragments of the *E. coli* gene, as a probe in a Southern analysis of whole cell DNA

from another organisms to identify the corresponding δ' homologue; (2) use of holA, or its fragments, as a probe to screen cDNA plasmid libraries of other organisms; (3) use of the holA gene sequence to synthesize oligonucleotide primers for PCR to amplify the corresponding δ' homologue from total genomic DNA from other organisms; and (4) use of the holA gene sequence to identify the δ' homologue from a genome sequencing project of other organisms by sequence comparison to the *E. coli* holA gene (O'Donnell Declaration ¶ 14).

The present application fully discusses the isolation and sequencing of the δ' and δ subunits and their encoding genes for the polymerase III holoenzyme. In view of the disclosure of these experimental procedures, the known structural and functional homology of the δ' and δ subunits proteins from various sources such as numerous other prokaryotes, and the present amendment limiting claims 5, 14, 54, and 59 to an isolated DNA molecule encoding a protein subunit of polymerase III holoenzyme from a eubacterial prokaryote and an isolated protein subunit of polymerase III holoenzyme from a eubacterial prokaryote, it would not require an undue amount of experimentation for one skilled in the art to isolate and sequence the claimed δ' and δ proteins (and their encoding gene) from eubacterial prokaryote sources other than *E. coli*.

The rejection of claims 14-16, 21, 24, 32, 35, 43, 46, 54, 57, and 66-75 under 35 U.S.C. § 112, first paragraph, for lack of written description is respectfully traversed in view of the above remarks. In addition, claims 21, 24, 32, 35, 43, 46, and 66-75 have been canceled.

The rejection of claims 66-75 under 35 U.S.C. § 112, second paragraph, for indefiniteness is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 32-35 under 35 U.S.C. § 102(b) as anticipated by Yoshikawa et al., "Cloning and Nucleotide Sequencing of the Genes *rimI* and *rimJ* which Encode Enzymes Acetylating Ribosomal Proteins S18 and S5 of *Escherichia coli* K12," Mol. Gen. Genet., 209:471-488 (1987) ("Yoshikawa") is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 36, 37, 39, 41, and 42 under 35 U.S.C. § 102(b) as anticipated by Yoshikawa is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 43, 45, and 46 under 35 U.S.C. § 102(b) as anticipated by Stirling et al., "*xerB*, an *Escherichia coli* Gene Required for Plasmid ColE1 Site-Specific Recombination, is Identical to *pepA*, Encoding Aminopeptidase A, a Protein with Substantial

Similarity to Bovine Lens Leucine Aminopeptidase," EMBO J., 8:1623-1627 (1989) ("Stirling") is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 47, 49, 50, 51, 52, and 53 under 35 U.S.C. § 102(b) as anticipated by Stirling is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 54 and 56-58 under 35 U.S.C. § 102(b) as anticipated by Takase et al., "Genes Encoding Two Lipoproteins in the *leuS-dacA* Region of the *Escherichia coli* Chromosome," J. Bac., 169:5692-5699 (1987) ("Takase") is respectfully traversed.

Takase relates to the coding of two lipoproteins by two genes, *rlpA* and *rlpB*, located in the *leu*S-*dac*A region on the *Escherichia coli* chromosome (O'Donnell Declaration ¶ 17). The *rlpA* gene encodes for a lipoprotein having molecular weight of 36K (Id.). Figure 6 of the reference details the sequence of the 36K lipoprotein gene *rlpA* and its 5'- and 3'-flanking regions and the amino acid sequences deduced from the nucleotide sequence (Id.). The position of the PTO is that this sequence matches that of the sequence encoding the claimed δ subunit. Applicants respectfully disagree. This sequence is not holA, it is rlpA and rlpB, the subject of Takase. At the end of the sequence, past both rplA and B, are 230 base pairs that was not discussed (Id.). This sequence encodes the first 20-25% of the holA gene sequence (Id.). Takase did not recognize this to be an open reading frame of a putative unknown gene, nor did the reference disclose the complete sequence of the holA gene (Id.). The diagram attached as Exhibit 9 to the O'Donnell Declaration shows the overlap between the disclosed *rlpB* gene of Takase and the holA gene encoding the claimed δ subunit (Id.). Thus, the δ protein subunit of polymerase III holoenzyme and the gene encoding the δ protein subunit of the polymerase III holoenzyme of the present invention are not disclosed by Takase.

In particular, Takase only discloses a portion of the holA gene encoding the δ protein subunit of polymerase III holoenzyme and does not disclose the nucleotide or protein sequences for the entire δ subunit.

In contrast, claim 54 relates to "[a]n isolated protein subunit of polymerase III holoenzyme from a eubacterial prokaryote, wherein the subunit group is δ." Further, claim 59 relates to "[a]n isolated DNA molecule encoding a protein subunit of polymerase III holoenzyme from a eubacterial prokaryote, wherein the subunit group is δ." Takase does not teach the *entire* specified isolated protein subunits of polymerase III holoenzyme, nor the *entire* gene encoding that protein. Further, Takase does not disclose the claimed expression

system or host cell. Since Takase does not disclose the entire δ protein subunit nor the entire sequence encoding the δ protein subunit, there is no basis for an anticipation rejection.

The outstanding office action places great reliance on the results from use of the MPSearch sequence analysis software employing the Smith-Waterman algorithm. Applicant has not been provided with this analysis or algorithm and, therefore, has great difficulty responding to this aspect of the outstanding office action. In any event, however, it is beyond dispute that Takase fails to disclose the complete sequences for the δ subunit. The MPSearch sequence analysis and Smith-Waterman algorithm are thus contrary to fact, as demonstrated by the O'Donnell Declaration and Takase itself. Since there is no reasonable basis for the rejection over Takase, that rejection must be withdrawn.

The rejection of claims 59, 60, 64, and 65 under 35 U.S.C. § 102(b) as anticipated by Takase is respectfully traversed in view of the remarks in the preceding paragraphs.

The rejection of claims 32-35 under 35 U.S.C. § 103(a) as being unpatentable over Yoshikawa is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 41 and 42 under 35 U.S.C. § 103(a) as being unpatentable over Yoshikawa is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 43-46 under 35 U.S.C. § 103(a) as being unpatentable over Stirling is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 52 and 53 under 35 U.S.C. § 103(a) as being unpatentable over Stirling is respectfully traversed in view of the cancellation of these claims.

The rejection of claims 54-58 under 35 U.S.C. § 103(a) as being unpatentable over Takase is respectfully traversed.

As stated above, Takase does not disclose the *entire* specified isolated δ protein subunit of polymerase III holoenzyme, nor the *entire* gene encoding that protein. In particular, Takase discloses only a short portion of the gene encoding the δ protein subunit. In addition, Takase provides no motivation to determine the sequence of the remainder of the gene. Specifically, Takase failed to identify the open reading frame of the gene for the δ protein subunit of polymerase III holoenzyme and, therefore, provides no motivation or suggestion to determine the remainder of the gene encoding the δ protein subunit. Further, the focus of Takase is on two genes, *rlpA* and *rlpB*, which are different from the gene encoding the δ protein subunit of polymerase III holoenzyme. As a result, Takase provides

no motivation with respect to determining the sequence of the gene encoding the δ protein subunit of polymerase III holoenzyme. Therefore, the rejection based on this reference is improper and should be withdrawn.

The rejection of claims 5, 7, 8, 10, 12, 13, 17, 21, 22, 25, 26, 28, 32, 36, 37, 39, 41, 43, 47, 48, 50, 53, 55, 60, 62, and 65 under 35 U.S.C. § 101 as claiming the same invention as that of prior U.S. Patent Application Serial No. 08/279,058, now U.S. Patent No. 5,668,004 to O'Donnell ("the '004 Patent") is respectfully traversed.

Claims 21, 22, 25, 26, 28, 32, 36, 37, 39, 41, 43, 47, 48, 50, and 53 have been canceled. Further, the claims of the '004 Patent are limited to subunits of DNA polymerase III from *Escherichia coli*. In contrast, the claims of the present application are not limited to *E. coli*. Since the scope of the claims in the '004 Patent and the present application are not identical, the rejection for same invention type double patenting is improper and should be withdrawn.

The rejection of claims 9, 13, 17, 21, 23, 32, 38, 41, 43, 47, 49, 53, 63, and 65 for obviousness-type double patenting as being unpatentable over the '004 Patent, to the extent those claims remain in the present application, is respectfully traversed in view of the terminal disclaimer filed herewith.

In view of all the foregoing, it is submitted that this case is in condition for allowance and such allowance is earnestly solicited.

Respectfully submitted,

Date: June 8, 1999

Michael L. Goldman
Registration No. 30,727
Attorney for Applicant

Nixon, Hargrave, Devans & Doyle LLP
Clinton Square, P. O. Box 1051
Rochester, New York 14603
Telephone: (716) 263-1304
Facsimile: (716) 263-1600

Certificate of Mailing - 37 CFR 1.8 (a)

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Assistant Commissioner for Patents, Washington, D.C. 20231. on the date below.

6/8/99 | Wendy L. Harrold
Date | Wendy L. Harrold

#16
CAN
6-16-99

PATENT
Docket No.: 19603/10214 (CRF D-1156C)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Michael O'Donnell	Examiner: E. Stole
Serial No.	:	08/828,323	Art Unit: ~~1652~~ 1653
Filed	:	March 28, 1997	
For	:	DNA POLYMERASE III HOLOENZYME	

SUPPLEMENTAL DECLARATION OF MICHAEL O'DONNELL UNDER 37 CFR § 1.132

Assistant Commissioner for Patents
Washington, D.C. 20231

Dear Sir:

I, MICHAEL O'DONNELL, pursuant to 37 CFR § 1.132, declare:

1. I received a B.S. degree in Biochemistry from the University of Portland, Portland, Oregon in 1975 and a Ph.D. degree in Biochemistry from the University of Michigan, Ann Arbor, Michigan in 1982. I was a Postdoctoral Fellow at Stanford University, Stanford, California from 1982 to 1986.

2. I am a Professor, Rockefeller University, New York, New York. In addition, I am an Investigator with Howard Hughes Medical Institute, Chevy Chase, Maryland.

3. I am the sole named inventor of the above-identified application.

4. I present this declaration to show (1) that proteins homologous to the δ' and δ subunits of DNA polymerase III holoenzyme are contained in eubacterial prokaryotes other than *E. coli* and (2) that, using the sequence information for the *E. coli* δ' and δ subunits in my present application, those skilled in the art could obtain these subunits from other eubacterial prokaryotes and, in fact, have done so.

5. Various genome projects for many different organisms have resulted in the gene sequences for various bacteria being publicly available on various web sites. As described more fully below, the amino acid sequences for the δ and δ' subunits for *E. coli*,

disclosed in my present application, were used, by myself and others, in a BLAST search program (Altschul, et al., "Basic Local Alignment Search Tool," J. Mol. Bio. 215:403-10 (1990)) to identify the presence of genes encoding these proteins in other eubacterial prokaryotes. As explained in the textbook Molecular Biology of the Gene (attached hereto as Appendix A), eubacterial (i.e. true bacteria) prokaryotes are a distinct kingdom separate from eukaryotes and archaebacteria and include: Aquificales (included *Aquifex aeolicus*), *Chlamydiales*, *Coprothermobacter*, *Cyanobacteria*, Green Sulfur bacteria (includes *Porphyromonal gingivalis* and *Chlorobium tepidum*), *Fibrobacter* group, *Firmicutes* (Gram positives including *Mycobacterium*, *Clostridium acetobutylicum*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Staphylococcus aureus*, *Bacillus subtilis*), *Flexistipes* group, *Fusobacteria*, Green non-sulfur bacteria, *Holophaga* group, *Nitrospira* group, *Planctomycetales*, *Proteobacteria* (includes the alpha subdivision (e.g. *Caulobacter crescentus*), the beta group (e.g. *Bordetella pertussis* and *Neisseria meningltidis*), the delta/epsilon subdivisions (e.g. *Campylobacter jejuni* and *Helicobacter pylori*), and the gamma subdivision (e.g. the *Enterobacteriaceae* that includes *Haemophilus influenzai*, *Yersinia pestis*, *Vibrio cholerai*, *Escherichia coli*, *Pasturella multocida*, *Pseudomonas aeruginosa*, *Salmonells typhi*, *Shewanella putrefaciens*), *Spirochaetales* (includes *Borrelia burgdorgeri*, *Treponema palladum*), *Synergistes* group, *Thermodesulfobacterium* group, *Thermotogales* (included *Thermotoga maritima*), *Thermus/Deinococcus* group (included *Thermus thermophilis* and *Deinococcus radiodurans*), and a variety of as yet unclassified bacteria. The results of these analyses are set forth below.

6. The sequence analysis of *Haemophilus influenze* is found at http://www.tigr.org/tdb/mdb/hidb/hidb.html. A copy of that web site listing is attached at Appendix B with the δ subunit encoding gene being identified as HI0923 and the δ' subunit encoding DNA molecule being identified as HI0455. This listing shows that the δ subunit encoding DNA molecule of *Haemophilus influenze* is 62.0% similar to the δ subunit encoding DNA molecule of *E. coli*. Likewise, the δ' subunit of *Haemophilus influenze* is shown to be 57.4% similar to the δ' subunit encoding DNA molecule of *E. coli*.

7. The genome of *Niceria gonorrhoeae* is found at the web site http://www.genome.on.edu. Search for the δ subunit amino acid sequence yields a contig. with a very high probability of 1.2×10^{-25}, contig. 188, while the δ' amino acid sequence yields a contig. of high probability of 1.2×10^{-14} #200. See Appendix C.

8. The genome for *Shewanella putrefaciens* is found on the TIGR BLAST server. A search for the δ subunit produced the high score of 1.1×10^{-54} for contig. gsp 230, while the search for δ' subunit produced the high score of 6.4×10^{-27} for contig. gsp 271. See Appendix D.

9. The genome for *Vibrio cholerae* is found at http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=v.cholerae. A search for the δ subunit produced the high score of 6.9×10^{-82} for contig. asm 937, while the search for δ' subunit produced the high score of 8.1×10^{-37} for contig. asm 894. See Appendix E.

10. The genomes for *Pseudomonas aeruginosa* (see Appendix F), *Salmonella typhi* (see Appendix G), and *Yersinia pestis* (see Appendix H) are found at http://www.ncbi.nlm.nih.gov/Blast/unfinished genomes. For these, the amino acid sequence of *E. coli* δ and δ' were used in BLAST searches. The high scores, given below, are all sufficiently significant to call the identified gene the one that performs the homologous function in *E. coli*:

Pseudomonas aeruginosa
δ - 7×10^{-34} contig. 52
δ' - 9×10^{-27} contig. 50

Salmonella typhi
δ - 1×10^{-161} contig. 1564
δ' - 8×10^{-10} contig. 870

Yersinia pestis
δ - 1×10^{-127} contig. 803
δ' - 9×10^{-98} contig. 51

11. Thus, for Gram negative bacteria such as *Haemophilus influenze*, *Niceria gonorrhoeae*, *Shewanella putrefaciens*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Salmonella typhi*, and *Yersinia pestis*, there is a high level of homology between the δ and δ' subunits of those bacteria and the δ and δ' subunits of *E. coli*.

12. For other eubacteria, there is significant homology between their δ' subunit and that of *E. coli*. In all eubacteria, the δ subunit can be identified starting with the *E. coli* δ subunit as comparison, but, since it is not as conserved as the δ' subunit, one must "walk" from one organism to another, as discussed in ¶ 23 below.

13. In Himmelreich et al., "Complete Sequence Analysis of the Genome of the Bacterium *Mycoplasma pneumoniae*," Nucleic Acids Research 24(22):4420-4449 (1996), the

δ' subunit of *Mycoplasma pneumoniae* is identified as being homologous to the δ' subunit of *E. coli* in Table 1 on page 4426. See Appendix I.

14. In Kunst et al., "The Complete Genome Sequence of the Gram-positive Bacterium *Bacillus subtilis*," Nature 390:249-256 (1997), the δ' subunit of *Bacillus subtilis* is identified as being homologous to the δ' subunit of *E. coli* in the table on page 248. See Appendix J.

15. The genome for *Streptococcus pyogenes* is found in the University of Oklahoma server (i.e. http://www.ncbi.nlm.nih.gov.BLAST/tigrbl.html). δ' produced the high score of 3.3×10^{-10} for contig. 218. See Appendix K.

16. The genome for *Enterococcus faecalis* is found on the TIGR BLAST search server. δ' produced the high score of 9.6×10^{-16} for contig. 6277. See Appendix L.

17. The genome for *Streptococcus pneumoniae* is found on the TIGR BLAST search server. δ' produced the high score of 2.4×10^{-12} for contig. sp 68. See Appendix M.

18. The genome for *Aquifex aeolicus* is found in Deckert et al., "The Complete Genome of the Hyperthermophilic bacterium *Aquifex aeolicus*," Nature 392:353-358 (1998) and at http://www.ncbi.nlm.nih.gov/Blast/unfinished genomes. δ' produced the high score of 8×10^{-13} (position 1303996-1304394). See Appendix N.

19. The genome for *Thermatoga maritima* is found in the TIGR BLAST server page. δ' yields a high score of 3.7×10^{-15} for contig. tm 26. See Appendix O.

20. In *Spirochaetes*, Tomb et al., "The Complete Genome Sequence of the Gastric Pathogen *Helicobacter pylori*," Nature 388:539-547 (1997) (see Appendix P) and Fraser et al., "Genomic Sequence of a Lyme Disease Spirochaete, *Borrelia burgdorferi*," Nature 390:580-586 (1997) (see Appendix Q), *Helicobacter pylori* and *Borrelia burgdorferi* are identified to have δ' subunits. For *Helicobacter pylori*, δ' is listed in the table as HP1231. For *Borrelia burgdorferi*, using the NCBI genome search page (Ncbi.nlm.nih.gov/Blast/unfinished genomes), δ' gives the high score of 8×10^{-7}. See Appendix R.

21. In Andersson et al., "The Genome Sequence of *Rickettsia prowazekii* and the Origin of Mitochondria," Nature 396:133-140 (1998), *Rickettsia prowazekii* is identified to have a δ' subunit, identified as RP172. See Appendix S.

22. A large compilation of genome sequences is at the web site http://www.ncbi.nlm.gov/Blast/unfinished genome.html. The eubacterial genomes were searched using the δ' subunit of *E. coli* . All organisms in eubacteria scored very high with

identity levels sufficient to identify the holB gene encoding δ' conclusively. This is seen in Figure 1 showing a path of one-on-one comparative alignments each of which start with *E. coli* and the alignments (attached hereto as Appendix T). In this figure, within the paratheses, is the percent identity and the ratio of the number of identities (i.e. the numerator) over the length of the amino acid sequence that was compared (i.e. the denominator). The number outside of the parentheses is the score obtained in the Blast program (i.e. even a score of 1 x 10^{-9} is a sufficiently high score to identify the homologous gene).

23. A similar search with the δ subunit of *E. coli* identified the holA gene of *Nisseria* and *Thiobacillus* as high matches, and holA of other enteric bacteria produced high scores as well. Repetition of this procedure using *Neisseria* δ easily allows the identification of δ in *Aquifex aeolicus*. Use of *Aquifex aeolicus* δ identifies δ of *Enterococcus* (which identifies *Bacillus* δ, then *Streptococcus* δ, then *Synechocystis*, and the *Porphyromonas* δ). Use of *Aquifex aeolicus* δ also identifies *Thermatoga* δ, which identifies Spirochaetes (*Borrelia*) δ subunit. Use of *Thiobacillus* δ identifies δ from *Helicobacter camylobacter*. There is a region at about 100 residues that is rather well conserved in δ across eubacteria and if this were used, the scores could be even higher yet. Figure 2 shows this "walking" procedure and shows the scores and percent identities obtained as a result of this procedure starting from the δ subunit of *E. coli* as well as alignments (attached hereto at Appendix U). This figure is substantially the same as Figure 1 but within the parentheses, after the percentage identity, there is another ratio and another percentage based on homologies. Figure 2 does not show scores for individual Gram negative bacteria of the *Enterobacteria* class (called enterics) as they are highly related to *E. coli* and the scores are very high.

24. As demonstrated by all the foregoing, those of ordinary skill in the art would have been able to (and, in fact, did) identify and isolate the δ and δ' subunits of their polymerases (and the encoding genes) from eubacteria other than *E. coli*.

25. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Date: 6/7/99

Michael E. O'Donnell

Michael E. O'Donnell

PATENT
19603/10212 (CRF D-1156A)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	Michael O'Donnell	Examiner: K. Hendricks
Serial No.:	08/279,058	Art Unit: 1814
Filed :	July 22, 1994	
For :	DNA POLYMERASE III HOLOENZYME	

RECEIVED
TECH CENTER 1600/2900
98 DEC -1 PM 12:02

DECLARATION OF MICHAEL O'DONNELL UNDER 37 CFR § 1.132

Assistant Commissioner for Patents
Washington, D.C. 20231

Dear Sir:

I, MICHAEL O'DONNELL, pursuant to 37 CFR § 1.132, declare:

1. I received a B.S. degree in Biochemistry from the University of Portland, Portland, Oregon in 1975 and a Ph.D. degree in Biochemistry from the University of Michigan, Ann Arbor, Michigan in 1982. I was a Postdoctoral Fellow at Stanford University, Stanford, California from 1982 to 1986.

2. I am a Professor of Molecular Biology, Department of Microbiology, Cornell University Medical Center, New York, New York.

3. I am the sole named inventor of the above-identified application.

4. In the following paragraphs, I: (1) describe why Maki et al, "DNA Polymerase III Holoenzyme of *Esherichia Coli*," J. Biol. Chem., 263(14):6547-54 (1988) ("Maki") does not disclose conformationally correct subunits of polymerase holoenzyme; (2) show that proteins homologous to the δ' subunit of polymerase III holoenzyme are contained in organisms other than *E. coli*; and (3) show that Takase et al., "Genes Encoding Two Lipoproteins in the *leu*S-*dac*A Region of

the *Escherichia coli* Chromosome," J. Bacteriology, 169(12):5692-99 (1987) ("Takase") does not disclose the δ subunit of polymerase III holoenzyme.

Maki Does Not Disclose Active Subunits

5. I performed my postdoctoral studies with Dr. Arthur Kornberg at Stanford, and worked in the same laboratory as Hisaji Maki, the first listed author of the Maki reference. Accordingly, I am knowledgeable regarding the work discussed in Maki. Although Figure 4 on page 6551 of Maki shows bands identified as the subunits of Polymerase III holoenzyme, there was uncertainty in the Kornberg laboratory as to the authenticity of the various bands. In particular, it was unclear whether or not these bands were true subunits. At the time, the only true and established subunits were the β, γ, τ, α, and ϵ proteins, as their genes mapped to classic temperature sensitive mutant alleles of DNA replication. However, no other classic temperature sensitive mutants in replication were left that had not already been identified. Hence, the bands shown in Figure 4 labelled δ, δ', χ, Ψ, and θ may have been either protein contaminants that were still present in the holoenzyme preparation or proteolytic products of the larger subunits (e.g. α, τ, γ). Indeed, most people in the field, did not believe that these protein bands were true subunits of the holoenzyme.

6. Further, I am familiar with the procedures described in Maki utilized to separate the subunits of the polymerase III holoenzyme. An important difference between Maki and my invention is that the proteins of my invention are purified without the use of denaturants. Maki discloses the use of a denaturant to separate the subunits, because they are tightly held into a particle of all ten proteins called the Pol III holoenzyme. Within this holoenzyme particle, there are 18 polypeptide chains, because some of the proteins are present in copies of two or more. Hence, to separate the subunits, Maki discloses the use of sodium dodecyl sulfate

("SDS") to denature the holoenzyme particle. SDS is one of the very most powerful protein denaturants, it completely unfolds polypeptide chains to form a rodlike SDS-polypeptide complex (Lehninger, A., Biochemistry, Worth Publishers, NY, NY, Third Edition, pp. 180 (1977)) (attached hereto as Exhibit 1). Samples for the SDS-PAGE technique, such as used by Maki et al., are typically boiled for 2-5 min. prior loading on the gel (See et al., "Estimating Molecular Weights of Polypeptides by SDS Gel Electrophoresis," In Protein Structure: A Practical Approach, IRL Press, New York ed. T.E. Creighton, pp. 1-21 (1989) (attached hereto as Exhibit 2)). The use of high temperatures and SDS will cause complete denaturation of most proteins. Id. Only in some cases is it possible to renature the proteins from an SDS-PAGE, and, then, it is often only useful for performing immunoprecipitations (Scheidtmann, K.H., "Immunological Detection of Proteins of Known Sequence," In Protein Structure: A Practical Approach, IRL press, New York, ed. T.E. Creighton, pp. 93-115 (1989) (attached hereto as Exhibit 3)). The basis for the antibody recognition of proteins lacking correct 3D conformation and full biological activity is that most antibodies recognize the primary sequence of the protein rather than requiring a correct three dimensional structure.

7. Once a protein is denatured in SDS, there is little hope of returning it to an active, or conformationally correct, form. I have tried this procedure with the δ, δ', and γ subunits without success in recovering activity. Generally, one must mince up the SDS gel, extract with a mortar and pestle, and remove the SDS using Dowex or acetone precipitation. Often, other denaturants such as urea and/or guanidine hydrochloride are used in the process. Guanidine hydrochloride and urea are polar molecules with no substantial aliphatic character and, therefore, can be efficiently dialyzed off a protein to permit renaturation in some cases. However, SDS has a large aliphatic component which binds tightly to protein and is difficult to remove completely, making renaturation unlikely.

8. In the examples of the present application, the subunits are all purified in the absence of denaturant and, accordingly, are conformationally correct throughout their purification. This is possible, because, in each case, an individual gene is used to make each isolated protein subunit. When only one subunit is produced in large amounts, the low intracellular level of the other 9 subunits is overwhelmed, and, thus, the single recombinant protein can be purified and recovered in isolation away from the other 9 protein subunits with which it would normally associate. Since interacting partner subunits are not present, denaturants are not needed to obtain the subunit in isolation from other subunits.

9. The activity of the proteins purified in accordance with the present application is demonstrated in the present specification by virtue of their being functional in a variety of assays including: (a) binding to other subunits, (b) activating or stimulating ATPase activity when in combination with other subunits, (c) activating or stimulating replication activity when in combination with other subunits, (d) activating or stimulating 3'-5'exonuclease activity when in combination with other subunits.

Proteins Homogolous to the δ' Subunit of Polymerase III Holoenzyme are Contained in Organisms Other than *E. Coli*

10. Most often, a protein active in *E. coli* has a homologous counterpart in higher cells. This is especially expected to be true of processes that are essential to life, such as DNA replication. Processes underlying other critical-to-life processes such as transcription, and ribosome-mediated translation, are also conserved in evolution. Some proteins in these processes are so similar in prokaryotes and eukaryotes that they can be exchanged for one another *in vivo*, and use of prokaryotic genes can lead to identification of the eukaryotic counterpart. All cells utilize DNA for their genetic material which must be duplicated to propagate the species. Hence, it can be anticipated that the central life

process of DNA duplication will also be conserved during evolution such that it will be performed similarly in prokaryotes and eukaryotes. In fact, prokaryotic replicase components are similar in structure and function to their eukaryotic counterparts and can substitute for the eukaryotic components in complex multiprotein replication systems involving numerous other proteins.

11. It was generally understood in the field, before the filing date of the present application, that mechanisms of replication are widely conserved in organisms spanning the evolutionary scale. Homology in structure and function of the replication apparatus from prokaryotes to eukaryotes had already been established from work of a variety of different laboratories. It was known that the bacteriophage T4 sliding clamp (gene 45 protein) was structurally homologous to the human PCNA clamp (Tsurimoto et al. "Functions of Replication Factor C and Proliferating Cell Nuclear Antigen: Functional Similarity of DNA Polymerase Accessory Proteins From Human Cells and Bacteriophage T4," PNAS, 87:1023-1027 (1990) ("Tsurimoto")). Moreover, it was known that the 3 components of the bacterial replicases (T4 and *E. coli*) are so homologous in structure and function to the human 3 component replicase, that the 3 components of these replicases (clamp, clamp loader, polymerase), could substitute in the place of the human 3 components in duplication of the SV40 chromosome with several other human replication proteins that these replicases need to work with (Matsumoto, et al, PNAS, 87:9712-26 (1990); Tsurimoto). In other words, the bacterial 3-component replicases were active with other human replication proteins that coordinate their actions with the replicase to duplicate the SV40 DNA genome (a eukaryotic virus). The other human proteins in these assays that the 3-component replicases of *E. coli* and phage T4 can work with are: the 3-subunit human RPA factor, the 4-subunit human priming machinery, the human topoisomerase, and the SV40 viral large T antigen. The fact that the bacterial replicases (phage T4 and *E. coli*) can work with these other replication

proteins shows that they must have very similar structures and that the points of contact between these proteins must be evolutionarily conserved at the level of the DNA sequence of the genes.

12. Furthermore, in Sanders et al., "Rules Governing the Efficiency and Polarity of Loading a Tracking Clamp Protein Onto DNA: Determinants of Enhancement in Bacteriophage T4 Late Transcription," The EMBO Journal 14(16):3966-76 (1995)("Sanders"), the common elements of structure and function of replicative DNA polymerases of eukaryotes, prokaryotes, and certain viruses are discussed. It is disclosed that the replicative DNA polymerases of all of these sources are composed of a core enzyme and a set of accessory proteins. Further, Stillman, "Smart Machines at the DNA Replication Fork," Cell 78:725-28 (1994) ("Stillman") discusses the functional similarity of proteins from *E. coli*, humans, and phage T4 that cause replication. Specifically, these exhibits show that *E. coli* contains an accessory complex called γ complex which includes the subunits γ, δ, δ', Ψ, and χ. Further, these exhibits show that homologous proteins to the γ complex are also present in eukaryotic (containing RFC complex), phage T4 (containing g44 complex), and human (containing RFC complex) organisms. Further, it was known that replicases of humans, *E. coli*, and the T4 virus were functionally, as well as structurally, homologous. They each have 3 components: (1) a DNA polymerase, (2) a processivity factor (sliding clamp), and (3) and a 5-protein ATPase that functions with the processivity factor to load it onto DNA.

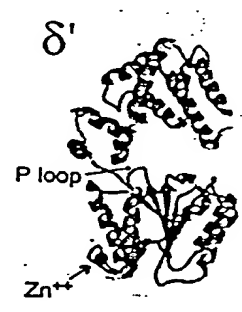
13. Furthermore, since *E. coli* and humans are at opposite ends of the evolutionary scale, it would have been known to one skilled in the art that all other bacteria and eukaryotes between *E. coli* and humans would also have structural homologues to the δ' subunit. Further, those skilled in the art recognize the δ' subunit from *E. coli* has sequence homology to accessory protein complexes of various other organisms. For example, in O'Donnell et al., "Homology

in Accessory Proteins of Replicative Polymerases - *E. coli* to Humans," Nucleic Acids Research 21(1):1-3 (1993) ("O'Donnell"), a comparison of amino acid sequences shows the homology between proteins of replicative polymerases of *E. coli*, humans, and phage T4. In Carter, et al., "Identification, Isolation, and Characterization of the Structural Gene Encoding the δ' Subunit of *Escherichia coli* DNA Polymerase III Holoenzyme," J. of Bacteriology, 175(12):3812-22 (1993), Figure 5 diagrams the homology of the δ' amino acid sequence to other replication proteins. Comparison of the δ' amino acid sequence revealed similarity to the A1(replication factor C) complex of HeLa cells and to the gene 44 protein (gp44) of bacteriophage T4. In addition, amino acid sequence similarity was found to the gene product of *B. subtilis*. Id. Further, the structural homology of the δ' subunit to other replication proteins has been proven to be true. Cullman, et al., "Characterization of the Five Replication Factor C Genes of *Saccharomyces cerevisiae*," Molecular and Cellular Biology, 15(9):4661-71 (1995). For example, the genome project of *Haemophilus influenze* showed homologues to all 10 subunits of *E. coli* DNA polymerase III holoenzyme, including δ, δ', χ, Ψ and θ. Currently, the GenBank now also shows homologues to the δ' subunit of *E. coli* from a large variety of organisms, including the following: Procaryotes: *Escherichia coli*, *Haemophilus influenze*, *Micrococcus luteus*, *Pseudomonas aeruginosa*, *Bacillus subtilis*, *Caulobacter crescentus;* Archaebacteria: *Thermus thermophilis* (extreme thermophile); Eukaryotes: *Drosophila melanogaster* (fly, insect), *Caenorhabditis elegans* (namatode, worm), *Gallus gallus* (dog), *Homo sapien* (man), *Saccharomyces cerevisiae* (yeast), and *Saccharomyces pombe* (yeast).

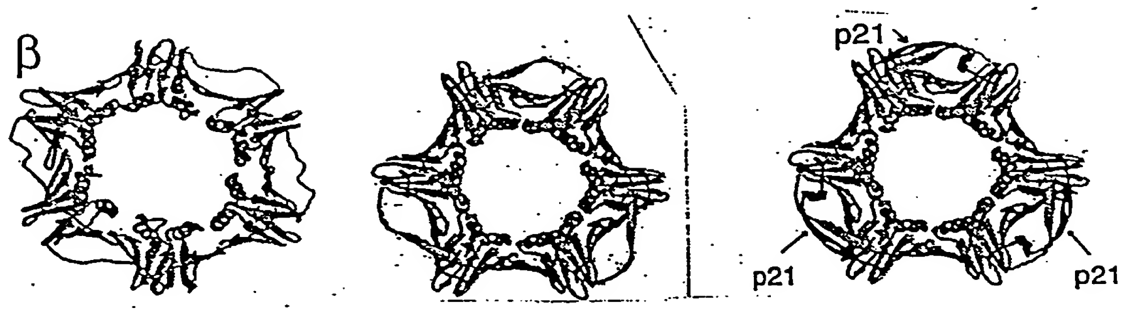
14. The sequence of the human homologues to δ', and indeed the other δ' homologues, are sufficiently homologous to the δ' subunit of *E. coli* to provide for identifying and obtaining the corresponding δ' (holA) gene from these organisms using the gene encoding the δ' subunit of *E. coli* in the following ways: (1) use of the *E. coli* holA gene, or

fragments of the *E. coli* gene, as a probe in a Southern analysis of whole cell DNA from another organisms to identify the corresponding δ′ homologue; (2) use of holA, or its fragments, as a probe to screen cDNA plasmid libraries of other organisms; (3) use of the holA gene sequence to synthesize oligonucleotide primers for PCR to amplify the corresponding δ′ homologue from total genomic DNA from other organisms; and (4) use of the holA gene sequence to identify the δ′ homologue from a genome sequencing project of other organisms by sequence comparison to the *E. coli* holA gene.

15. I have solved the structure of the δ′ protein (in collaboration with Dr. John Kuriyan's laboratory at Rockefeller University). The δ′ protein is composed of three domains in the shape of a C, and likely performs the clamp loading action by relative motions between the top and bottom domains allowing it to open and close the sliding clamp ring around DNA. The homology of E. coli δ′ to the δ′ of the several homologues listed above in paragraph 13 is well above the level needed to predict that they will have the exact same chain fold and C-shape.



16. The crystal structure of E. coli β clamp, the T4 gp45 clamp, and PCNA have been solved by my lab (in collaboration with Kuriyan's lab at Rockefeller University). They have the same chain fold and three dimensional structure (see below) (β subunit is shown in Kong et al., Cell, 69:425-37 (1992); yeast PCNA is in Krishna, et al., Cell, 79:1233-43 (1994); human PCNA is unpublished, T4 PCNA is unpublished).



Takase Does Not Disclose the δ Subunit of Polymerase III Holoenzyme

17. Takase relates to the coding of two lipoproteins by two genes, *rlpA* and *rlpB*, located in the *leuS-dacA* region on the *Escherichia coli* chromosome. The *rlpA* gene encodes for a lipoprotein having molecular weight of 36K. Figure 6 of the reference details the sequence of the 36K lipoprotein gene *rlpA* and its 5′- and 3′- flanking regions and the amino acid sequences deduced from the nucleotide sequence. Figure 7 of the reference details the sequence of the *rlpB* gene. At the end of the sequence in Figure 7, the last 230 base pairs constitute a sequence that encodes the first 20-25% of the holA sequence. Takase did not recognize this to be an open reading frame of a putative unknown gene, nor did this reference disclose the gene. See the diagram attached hereto as Exhibit 4. Further, as shown in Dong, et al., "DNA Polymerase III Accessory Proteins," J. Biological Chem., 268(16):11758-765, 11759 n. 3 ("Dong"), Takase's published sequence was incorrect and incomplete, in fact, the first 54 nucleotides of the δ gene are incorrect by 11 nucleotides. Thus, the δ protein subunit of polymerase III holoenzyme and the gene encoding the δ protein subunit of the polymerase III holoenzyme of the present invention are not disclosed by Takase.

18. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

12/6/96
Date

Michael O'Donnell
Michael O'Donnell

COMPLETE

MOLECULAR BIOLOGY OF THE GENE

FOURTH EDITION

James D. Watson — COLD SPRING HARBOR LABORATORY

Nancy H. Hopkins — MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Jeffrey W. Roberts — CORNELL UNIVERSITY

Joan Argetsinger Steitz — YALE UNIVERSITY

Alan M. Weiner — YALE UNIVERSITY

The Benjamin/Cummings Publishing Company, Inc.

Menlo Park, California • Reading, Massachusetts • Don Mills, Ontario
Wokingham, U.K. • Amsterdam • Sydney • Singapore
Tokyo • Madrid • Bogota • Santiago • San Juan

Cover art is a computer-generated image of DNA interacting with the Cro repressor protein of bacteriophage λ. The image was prepared by the Graphic Systems Research Group at the IBM U.K. Scientific Centre, using coordinates courtesy of Brian W. Matthews, University of Oregon.

Editor: Jane Reece Gillen
Production Supervisor: Karen K. Gulliver
Editorial Production Supervisor: Betsy Dilernia
Cover and Interior Designer: Gary A. Head
Contributing Designers: Detta Penna, Michael Rogondino
Copy Editor: Janet Greenblatt
Art Coordinator: Pat Waldo
Art Director and Principal Artist: Georg Klatt
Contributing Artists: Joan Carol, Cyndie Clark-Huegel, Barbara Cousins, Cecile Duray-Bito, Jack Tandy, Carol Verbeek, John and Judy Waller

Library of Congress Cataloging-in-Publication Data
Molecular biology of the gene / James D. Watson ...... [et al.].—4th ed.
p. cm.
Includes bibliographies and indexes.
ISBN 0-8053-9614-4
1. Molecular genetics. 2. Molecular biology. I. Watson, James D., 1928– .
QH447.M65 1988
574.87'328—dc19 88-4115
CIP

CDEFGHIJ-MU-943210

The Benjamin/Cummings Publishing Company, Inc.
2727 Sand Hill Road
Menlo Park, California 94025

as possible, with special emphasis on unusual bacteria that had previously eluded reliable phylogenetic placement.

Woese chose 16S rRNA for construction of a phylogenetic tree because it is truly universal and is so highly conserved in structure and function that phylogenetic trees are relatively easy to construct. In addition, 16S rRNA is an abundant RNA that can be quickly purified and analyzed even from small samples of cells. The early stages of 16S rRNA sequence analysis culminated in 1977 with a radical hypothesis. Woese proposed that procaryotes should be divided into two groups, called the archaebacteria and the eubacteria, which are as different from each other as either is from the eucaryotes. The clear implication of this proposal was that archaebacteria, eubacteria, and eucaryotes had all descended from an earlier common ancestor that did not survive. This hypothesis met with substantial resistance within the biological community because it contradicted two common but unfounded assumptions—that all bacteria are closely related and that bacteria more closely resemble the first living cells than do any eucaryotes.

Archaebacteria Assume Their Rightful Place[61-65]

Despite widespread scepticism about the value of dividing procaryotes into eubacteria and archaebacteria, proponents of the hypothesis continued to refine the universal phylogeny based on 16S rRNA (Figure 28-24) and to amass supporting biochemical evidence (see Table 28-5). Today, there is no longer any doubt that all living organisms belong to three coequal kingdoms, or **lines of descent**, and that none of these three kingdoms can be thought of as having given rise to the others (see Figure 28-24 and Table 28-5). Instead, all three have descended from an earlier living organism, or progenote, whose nature we can only infer by asking what archaebacteria, eubacteria, and the eucaryotic nucleus have in common. (The eucaryotic nucleus is directly descended from the progenote, but as we shall see, eucaryotic organelles such as the mitochondrion and chloroplast were derived by endosymbiosis of oxygen-fixing and photosynthetic eubacteria.)

The universal phylogeny based on 16S-like rRNA reveals other startling conclusions. Human beings (*Homo sapiens*) are in fact more closely related to corn (*Zea mays*) than a Gram-negative bacterium (*E. coli*) is to a Gram-positive bacterium (*Bacillus subtilis*) (see Figure 4-8 for the significance of Gram staining). Thus, the evolutionary distance separating two different bacteria can be greater than the distance between a sophisticated plant and the most sophisticated animal. The 16S-like phylogeny also provides definitive evidence for the endosymbiont hypothesis that mitochondria and chloroplasts are descended from eubacteria (see the section entitled The Endosymbiotic Origin of Mitochondria and Chloroplasts).

The Progenote (First Cell) Differed from All Modern Cells

The universal phylogeny based on 16S-like rRNA tells us that the three great kingdoms of living organisms are all descended from a progenote. But what was this progenote like? The abundance of

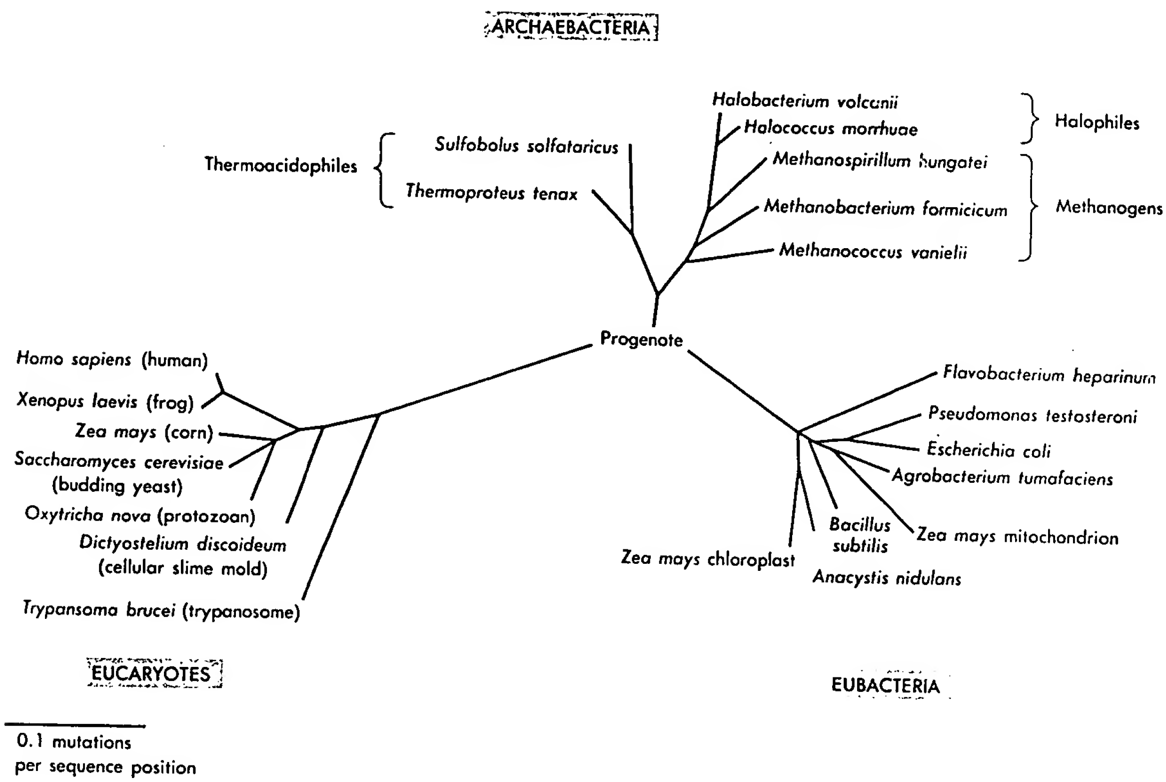


Figure 28-24
An evolutionary tree can be constructed by comparing the complete sequences of 21 different 16S and 16S-like ribosomal RNAs (rRNAs). The scale bar represents the number of accumulated nucleotide differences per sequence position in the rRNAs of the various organisms. Note that the scale bar cannot be recalibrated in billions of years without making the unjustified assumption that mutations accumulate in the DNA of all organisms at the same rate per unit time. [After N. R. Pace, G. J. Olsen, and C. R. Woese, *Cell* 45 (1986):325.]

introns in archaebacteria and eucaryotes suggests that the progenote had introns but that these were lost during eubacterial evolution as the genome was streamlined for very rapid growth (see Table 28-5). Similarly, since eubacteria and eucaryotes have ester-linked unbranched lipids containing L-glycerophosphate, it is likely that the progenote did, too.

Table 28-5 A Few of the Known Differences Between Archaebacteria and Eubacteria

Archaebacteria	Eubacteria
Genomic rearrangements common	Genomes quite stable
Transposable elements often abundant	Few transposable elements
Some introns in rRNA and tRNA	No introns known
No peptidoglycan in cell wall	Peptidoglycan cell wall
Branched-chain fatty acids	Straight-chain fatty acids
Ether-linked lipids	Ester-linked lipids
Lipids contain D-glycerophosphate	Lipids contain L-glycerophosphate
rRNA, tRNA, and ribosomes share both eubacterial and eucaryotic features	
Larger multisubunit RNA polymerases resembling the eucaryotic enzymes	Simpler RNA polymerases
"Reverse" gyrases in thermophiles introduce + supercoils	Gyrases introduce only − supercoils
EF2 sensitive to diphtheria toxin as in eucaryotes	EF2 insensitive to diphtheria toxin

But we cannot automatically assume that any trait shared by two of the three great kingdoms must reflect the nature of the progenote. Such a shared trait could also have arisen more than once as different organisms independently discovered its value. Independent evolution of the same characteristic in separate branches of a phylogenetic tree is called **convergent evolution.**

Bacteria Are More Highly Evolved than Higher Organisms

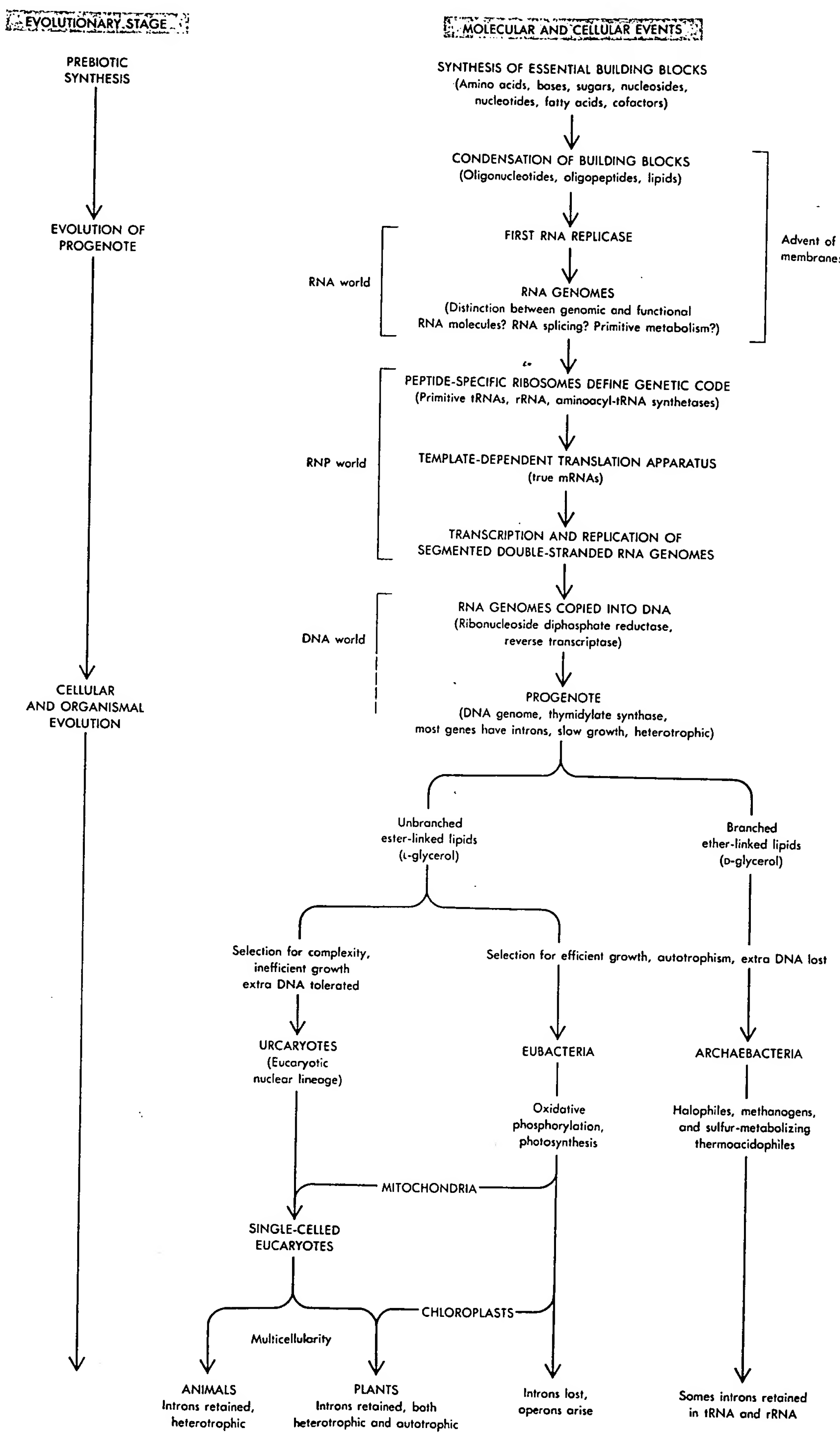
Efforts to deduce the nature of the progenote are also confounded by the fact that different organisms evolve at different rates. Although mutations arise in DNA throughout the life cycle of an organism, the effect of these mutations on fitness can only be tested in each *new* generation. As a result, rapidly multiplying organisms like bacteria and many lower eucaryotes have had a far greater opportunity to lose or modify the characteristics of the progenote than have more slowly growing higher organisms. This implies that many bacteria, although they are no more ancient than eucaryotes (see Figure 28-24), are actually more highly evolved.

The Endosymbiotic Origin of Mitochondria and Chloroplasts[66–70]

Eucaryotic cells contain a variety of internal organelles, each surrounded by a lipid bilayer. Many of these organelles (e.g., lysosomes, peroxisomes, and the endoplasmic reticulum) are relatively simple (see Figure 18-8). But two of them, mitochondria and chloroplasts, are about the same size as bacteria and, like bacteria, have circular DNA genomes (see Figure 15-17). Mitochondrial and chloroplast genomes encode the rRNA and tRNA components of the organellar translation apparatus, as well as mRNAs for organellar proteins that are synthesized within the organelle. The mitochondrial and chloroplast ribosomes are sensitive to antibiotics such as chloramphenicol, which kill many bacteria but do not affect the cytoplasmic ribosomes of eucaryotes.

The resemblance of mitochondria and chloroplasts to bacteria naturally led to the idea that these organelles began as free-living bacteria that had been engulfed by a primitive eucaryote (the **urcaryote**; see Figure 28-25). Once internalized, these symbiotic bacteria flourished within the host eucaryote as **endosymbionts**, while supplying the host with the ability to generate energy by oxidative phosphorylation and (in the case of plants) by photosynthesis. As the protomitochondrion and the protochloroplast slowly degenerated into specialized organelles, genes were transferred from organellar DNA to the nuclear genome of the host, leaving only a handful of essential genes behind in the organelle. As a result, most mitochondrial and chloroplast proteins are now encoded in the nuclear DNA, translated in the cytoplasm, and transported across the outer

Figure 28-25
A possible scheme for early evolution. [After J. E. Darnell and W. F. Doolittle, *Proc. Nat. Acad. Sci.* 83 (1986):1271.]


EVOLUTIONARY STAGE
PREBIOTIC SYNTHESIS
EVOLUTION OF PROGENOTE
CELLULAR AND ORGANISMAL EVOLUTION
MOLECULAR AND CELLULAR EVENTS
SYNTHESIS OF ESSENTIAL BUILDING BLOCKS
(Amino acids, bases, sugars, nucleosides, nucleotides, fatty acids, cofactors)
CONDENSATION OF BUILDING BLOCKS
(Oligonucleotides, oligopeptides, lipids)
Advent of membranes
RNA world
FIRST RNA REPLICASE
RNA GENOMES
(Distinction between genomic and functional RNA molecules? RNA splicing? Primitive metabolism?)
RNP world
PEPTIDE-SPECIFIC RIBOSOMES DEFINE GENETIC CODE
(Primitive tRNAs, rRNA, aminoacyl-tRNA synthetases)
TEMPLATE-DEPENDENT TRANSLATION APPARATUS
(true mRNAs)
TRANSCRIPTION AND REPLICATION OF SEGMENTED DOUBLE-STRANDED RNA GENOMES
DNA world
RNA GENOMES COPIED INTO DNA
(Ribonucleoside diphosphate reductase, reverse transcriptase)
PROGENOTE
(DNA genome, thymidylate synthase, most genes have introns, slow growth, heterotrophic)
Unbranched ester-linked lipids (L-glycerol)
Branched ether-linked lipids (D-glycerol)
Selection for complexity, inefficient growth extra DNA tolerated
Selection for efficient growth, autotrophism, extra DNA lost
URCARYOTES
(Eucaryotic nuclear lineage)
EUBACTERIA
ARCHAEBACTERIA
Oxidative phosphorylation, photosynthesis
Halophiles, methanogens, and sulfur-metabolizing thermoacidophiles
MITOCHONDRIA
SINGLE-CELLED EUCARYOTES
CHLOROPLASTS
Multicellularity
ANIMALS
Introns retained, heterotrophic
PLANTS
Introns retained, both heterotrophic and autotrophic
Introns lost, operons arise
Somes introns retained in tRNA and rRNA

membrane of the organelle. Only those molecular species that cannot cross the outer membrane (rRNA, tRNA, mRNA, and some proteins) must still be encoded in the organellar genome. (Some RNA molecules *can* cross the organellar membrane, however, as shown by the recent discovery that the RNA component of a mammalian mitochondrial RNA processing enzyme resembling RNase P is encoded by a nuclear gene.) As fewer and fewer proteins were encoded within the evolving endosymbiont, the organellar translation system no longer had to be extremely accurate. Eventually, the organellar translation apparatus degenerated into an apparently minimal translation machine (see Figure 14-15) and the mitochondrial genetic code underwent some surprising changes (see Table 15-9).

Did mitochondria and chloroplasts evolve from eubacterial or archaebacterial progenitors? Comparison of bacterial and eucaryotic cytochrome *c* initially suggested that mitochondria might have descended from the purple photosynthetic eubacteria. Comparison of animal mitochondrial and eubacterial 16S rRNA sequences failed to prove this, however, because the animal mitochondrial rRNA sequences had diverged too extensively to permit a meaningful comparison. Fortunately, plant mitochondrial 16S rRNAs are less divergent, and in this case, the comparison led to a surprising result. Plant mitochondria descended from a group of purple eubacteria that includes rhizobacteria (see Figure 22-49), agrobacteria (see Figure 22-48), and rickettsias (see page 544). Even today, each of these procaryotes is able to live within or in very close association with eucaryotic cells. This makes the endosymbiont hypothesis all the more plausible and allows us to complete a tentative scheme for early evolution (Figure 28-25).

Several Bacteriophage T4 Genes Contain Self-Splicing Introns[71,72]

No introns have ever been found in *E. coli*, the most intensively studied of all procaryotes. The discovery in 1984 of an intron in a bacteriophage of *E. coli* therefore came as quite a shock in 1984. Three different bacteriophage T4 genes are now known to have self-splicing introns resembling the *Tetrahymena* rRNA intron. Two of the genes encode enzymes that convert RNA precursors into DNA precursors (thymidylate synthase and the small subunit of ribonucleoside diphosphate reductase; see Figure 28-20). By expressing high levels of these two enzymes, T4 diverts the metabolic resources of the infected bacterium from making RNA to making DNA, thereby increasing the yield of DNA-containing progeny phage.

Why do self-splicing introns interrupt useful T4 genes? Although it is possible that the introns are harmless or hard to get rid of, another fascinating possibility is that the introns might actually contribute to efficient phage growth. Recall that self-splicing is initiated by attack of a free guanine nucleotide on the 5′ splice site of the intron (see Figures 28-8 and 28-9). When high levels of guanine nucleotides are present early in infection, efficient self-splicing of the transcripts will produce mRNAs whose protein products catalyze conversion of RNA precursors into DNA precursors. As guanosine nucleotide levels begin to fall later in infection, the efficiency of self-splicing will decrease, and the rate of conversion of RNA precursors to DNA precursors will consequently slow down. T4 thus appears to use the de-

Haemophilus influenzae Genome

Role Report: DNA metabolism: DNA replication, recombination, and repair

Match Acc# is linked to the primary accession for the sequence used to make the putative identification. %Sim (percentage of similarity) links to an alignment of that accession with the predicted gene.

DNA replication, recombination, and repair

HI#	Putative Identification	Match Acc#	%Sim
HI0759	A/G-specific adenine glycosylase (mutY) {Escherichia coli}	EGAD:19423	75.1
HI1740	ATP-dependent DNA helicase (recG) {Escherichia coli}	EGAD:16079	80.7
HI0728	ATP-dependent DNA helicase (recQ) {Escherichia coli}	EGAD:24475	78.4
HI0649	ATP-dependent DNA helicase (rep) {Escherichia coli}	EGAD:20036	82.8
HI0387	ATP-dependent helicase (dinG) {Escherichia coli}	EGAD:90681	76.2
HI0993	chromosomal replication initiator protein (dnaA) {Escherichia coli}	EGAD:23918	80.6
HI0314	crossover junction endodeoxyribonuclease (ruvC) {Escherichia coli}	EGAD:13020	88.3
HI0209	DNA adenine methylase (dam) {Escherichia coli}	EGAD:16548	71.4
HI1264	DNA gyrase, subunit A (gyrA) {Escherichia coli}	EGAD:20471	85.0
HI0567	DNA gyrase, subunit B (gyrB) {Escherichia coli}	EGAD:21268	86.1
HI1188	DNA helicase II (uvrD) {Haemophilus influenzae}	EGAD:18013	97.7
HI1100	DNA ligase (lig) {Escherichia coli}	EGAD:14652	79.9
HI0403	DNA mismatch repair protein (mutH) {Escherichia coli}	EGAD:21368	81.1
HI0067	DNA mismatch repair protein (mutL) {Escherichia coli}	EGAD:8885	67.3
HI0707	DNA mismatch repair protein (mutS) {Escherichia coli}	EGAD:20341	84.0
HI0856	DNA polymerase I (polA) {Escherichia coli}	EGAD:23009	77.0

HI0739	DNA polymerase III, alpha subunit (dnaE) {Escherichia coli}	EGAD:20624	85.8
HI0992	DNA polymerase III, beta subunit (dnaN) {Escherichia coli}	EGAD:16056	80.3
HI1397	DNA polymerase III, chi subunit (holC) {Haemophilus influenzae}	EGAD:7641	69.8
HI0923	DNA polymerase III, delta subunit (holA) {Escherichia coli}	EGAD:15752	62.0
HI0455	DNA polymerase III, delta' subunit (holB) {Escherichia coli}	EGAD:20293	57.4
HI0137	DNA polymerase III, epsilon subunit (dnaQ) {Escherichia coli}	EGAD:21587	76.5
HI0011	DNA polymerase III, psi subunit (holD) {Escherichia coli}	EGAD:10576	59.1
HI1229	DNA polymerase III, subunits gamma and tau (dnaX) {Escherichia coli}	EGAD:21779	69.8
HI0532	DNA primase (dnaG) {Escherichia coli}	EGAD:12957	74.0
HI1597	DNA repair protein (radA) {Escherichia coli}	EGAD:5672	92.2
HI0952	DNA repair protein (radC) {Escherichia coli}	EGAD:24268	71.7
HI0070	DNA repair protein (recN) {Escherichia coli}	EGAD:7996	67.8
HI0332	DNA repair protein (recO) {Escherichia coli}	EGAD:14344	76.5
HI1365	DNA topoisomerase I (topA) {Escherichia coli}	EGAD:15040	84.4
HI0444	DNA topoisomerase III (topB) {Escherichia coli}	EGAD:20019	79.4
HI0654	DNA-3-methyladenine glycosidase I (tagI) {Escherichia coli}	EGAD:6433	76.0
HI0991	DNA/ATP binding protein (recF) {Escherichia coli}	EGAD:21167	76.1
HI0062	dnaK suppressor protein (dksA) {Escherichia coli}	EGAD:19456	85.2
HI1689	endonuclease III (nth) {Escherichia coli}	EGAD:18451	91.9
HI0249	excinuclease ABC, subunit A (uvrA) {Escherichia coli}	EGAD:6128	91.2
HI1247	excinuclease ABC, subunit B (uvrB) {Escherichia coli}	EGAD:9811	87.7
HI0057	excinuclease ABC, subunit C (uvrC) {Escherichia coli}	EGAD:10876	80.1
HI0041	exodeoxyribonuclease III (xthA) {Escherichia coli}	EGAD:16058	83.5
HI1322	exodeoxyribonuclease V, alpha chain (recD) {Escherichia coli}	EGAD:7587	59.3
HI1321	exodeoxyribonuclease V, beta chain (recB) {Escherichia coli}	EGAD:9606	58.2
HI0942	exodeoxyribonuclease V, gamma chain (recC) {Escherichia coli}	EGAD:8221	61.2

HI0946	formamidopyrimidine-DNA glycosylase (fpg) {Escherichia coli}	EGAD:8852	74.7
HI0582	glucose inhibited division protein (gidA) {Escherichia coli}	EGAD:14924	87.4
HI0486	glucose-inhibited division protein (gidB) {Escherichia coli}	EGAD:24336	78.0
HI0980	Hin recombinational enhancer binding protein (fis) {Escherichia coli}	EGAD:6539	92.9
HI0313	Holliday junction DNA helicase (ruvA) {Escherichia coli}	EGAD:24056	79.9
HI0312	Holliday junction DNA helicase (ruvB) {Escherichia coli}	EGAD:15420	90.3
HI1546	impA protein, putative {Escherichia coli}	EGAD:33192	53.7
HI0676	integrase/recombinase (xerC) {Escherichia coli}	EGAD:21922	75.4
HI0309	integrase/recombinase (xerD) {Escherichia coli}	EGAD:24225	84.8
HI1424	integrase/recombinase, putative {Escherichia coli}	EGAD:8989	55.5
HI1572	integrase/recombinase, putative, authentic point mutation {Escherichia coli}	EGAD:14856	57.0
HI1313	integration host factor, alpha-subunit (himA) {Escherichia coli}	EGAD:13310	83.0
HI1221	integration host factor, beta-subunit (himD) {Escherichia coli}	EGAD:10401	78.3
HI0402	methylated-DNA--protein-cysteine methyltransferase (dat1) {Bacillus subtilis}	EGAD:8823	61.7
HI0910	mutator mutT protein (mutT) {Escherichia coli}	EGAD:16514	72.0
HI0339	primosomal protein N' (priA) {Escherichia coli}	EGAD:19358	70.2
HI0546	primosomal replication protein N (priB) {Escherichia coli}	EGAD:22393	75.2
HI0600	recA protein (recA) {Haemophilus influenzae}	EGAD:13744	100.0
HI0443	recombination protein (recR) {Escherichia coli}	EGAD:19714	88.4
HI0599	regulatory protein (recX) {Pseudomonas fluorescens}	EGAD:10671	52.9
HI1574	replicative DNA helicase (dnaB) {Escherichia coli}	EGAD:5615	82.8
HI0138	ribonuclease H (rnh) {Escherichia coli}	EGAD:29509	76.8
HI0192	seqA protein (seqA) {Escherichia coli}	EGAD:7875	71.8
HI0250	single-stranded DNA binding protein (ssb) {Haemophilus influenzae}	EGAD:12943	98.2
HI0624	sun protein (sun) {Echerichia coli}	EGAD:36279	71.2
HI1529	topoisomerase IV, subunit A (parC) {Escherichia coli}	EGAD:9582	85.4
HI1528	topoisomerase IV, subunit B (parE) {Escherichia coli}	EGAD:10527	88.6

HI1258	transcription-repair coupling factor (mfd) {Escherichia coli}	EGAD:8856	83.0
HI0018	uracil DNA glycosylase (ung) {Escherichia coli}	EGAD:13350	80.0

OU Neisseria Gonorrhoeae Sequence Blast Server Results

```
TBLASTN 1.3.9 [29-Oct-93]
Reference:  Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers,
and David J. Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol.
215:403-410.
Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.
Query=  deltaprime.ecoli
        (334 letters)
Database:  /gono/abi/Gcphrap/auto_gono
           114 sequences; 2,133,469 total letters.
Searching..................................................done
                                                                Smallest
                                                                Poisson
                                                  Reading  High Probability
Sequences producing High-scoring Segment Pairs:     Frame Score  P(N)      N
Contig200                                             -3   147  1.2e-14    2
Contig189                                             -1   152  2.5e-14    1
Contig138                                             -1    95  2.7e-06    1
Contig190                                             +1    46  0.033      7
Contig199                                             -3    51  0.16       2
Contig201                                             -2    54  0.80       1
Contig188                                             +3    45  0.84       3
Contig187                                             +3    53  0.89       1
Contig181                                             +3    52  0.95       1
Contig176                                             +1    51  0.99       1
Contig146                                             +1    51  0.99       1
Contig191                                             -3    46  0.993      3
>Contig200
          Length = 93,974
  Minus Strand HSPs:
 Score = 147 (68.2 bits), Expect = 1.3e-13, P = 1.3e-13
 Identities = 35/98 (35%), Positives = 48/98 (48%), Frame = -3
Query:    62 CRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLT 121
             C  CQ            Y  L        +  G+D +REV E          G  KV   + +  +L+
Sbjct: 72738 CGVCQSCTQIDAGRYVDLLEIDAASNTGIDNIREVLENAQYAPTAGKYKVYIIDEVHMLS 72559
Query:   122 DAAANALLKTLEEPPAETWFFLATREPERLLATLRSRC 159
              +A NA+LKTLEEPP     F LAT +P ++   T+ SRC
Sbjct: 72558 KSAFNAMLKTLEEPPEHVKFILATTDPHKVPVTVLSRC 72445
 Score = 98 (45.4 bits), Expect = 1.2e-14, Poisson P(2) = 1.2e-14
 Identities = 19/55 (34%), Positives = 29/55 (52%), Frame = -3
Query:    21 GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPD 75
             GR HHA L+     G+G     +     L++ L C+  Q   + CG C+ C   + AG + D
Sbjct: 72852 GRLHHAYLLTGTRGVGKTTIARILAKSLNCENAQHGEPCGVCQSCTQIDAGRYVD 72688
 Score = 44 (20.4 bits), Expect = 13., Poisson P(2) = 1.0
 Identities = 8/14 (57%), Positives = 8/14 (57%), Frame = -2
Query:   238 HEQAPARLHWLATL 251
             H    P R HWLA L
Sbjct: 88705 HTPYPQRAHWLALL 88664
 Score = 44 (20.4 bits), Expect = 2.8, Poisson P(3) = 0.94
 Identities = 10/20 (50%), Positives = 13/20 (65%), Frame = -2
Query:   315 LLLRIEHYLQPGVVLPVPHL 334
```

```
             LL  +  YL+ GV+ PVP L
Sbjct: 10810 LLGMVARYLKLGVLKPVPSL 10751
 Score = 43 (19.9 bits), Expect = 1.1, Poisson P(4) = 0.67
 Identities = 8/24 (33%), Positives = 13/24 (54%), Frame = -3
Query:   137 AETWFFLATREPERLLATLRSRCR 160
             A+ W  + T+ P+  LA   + CR
Sbjct: 49281 AQWWLVICTQSPKIGLAMANAACR 49210
 Score = 41 (19.0 bits), Expect = 2.6, Poisson P(5) = 0.93
 Identities = 9/14 (64%), Positives = 9/14 (64%), Frame = -1
Query:   186 ALLAALRLSAGSPG 199
             A L ALR  AG PG
Sbjct: 15380 AFLQALRKGAGQPG 15339
>Contig189
          Length = 45,334
  Minus Strand HSPs:
 Score = 152 (70.5 bits), Expect = 2.5e-14, P = 2.5e-14
 Identities = 32/87 (36%), Positives = 51/87 (58%), Frame = -1
Query:    90 VDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWFFLATREPE 149
             +DAVRE+ + +    +  GG +V+ +  A  +   AAN+LLK LEEPP +  F L +   +
Sbjct: 29077 IDAVREIIDNVYLTSVRGGLRVILIHPAESMNVQAANSLLKVLEEPPPQVVFLLVSHAAD 28898
Query:   150 RLLATLRSRCRLHYLAGPPEQYAVTWL 176
             ++L T++SRCR   L  P    A+ +L
Sbjct: 28897 KVLPTIKSRCRKMVLPAPSHGEALAYL 28817
 Score = 86 (39.9 bits), Expect = 1.4e-11, Poisson P(2) = 1.4e-11
 Identities = 13/27 (48%), Positives = 16/27 (59%), Frame = -1
Query:    55 GHKSCGHCRGCQLMQAGTHPDYYTLAP 81
             G K CG C  C L   G+HPD+Y + P
Sbjct: 29203 GCKPCGECMSCHLFGRGSHPDFYEITP 29123
 Score = 45 (20.9 bits), Expect = 26., P = 1.0
 Identities = 8/20 (40%), Positives = 11/20 (55%), Frame = -1
Query:    44 LSRYLLCQQPQGHKSCGHCR 63
             LSR++    +P G   C  CR
Sbjct: 25822 LSRHISFNRPSGRFGCSGCR 25763
 Score = 43 (19.9 bits), Expect = 1.9, Poisson P(3) = 0.85
 Identities = 12/46 (26%), Positives = 20/46 (43%), Frame = -1
Query:   134 EPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSRE 179
             E P+E  F   T   +  +  L++   L  L   P +Y   W+ R+
Sbjct: 18817 EMPSENHFT*QTD*QKTRMTLLKNDTFLRALLKQPVEYTPIWMMRQ 18680
>Contig138
          Length = 6169
  Minus Strand HSPs:
 Score = 95 (44.1 bits), Expect = 2.7e-06, P = 2.7e-06
 Identities = 16/37 (43%), Positives = 25/37 (67%), Frame = -1
Query:     4 YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDAL 40
             YPWL P + ++  ++   G GHHA+LI+A  G+G + L
Sbjct:  1849 YPWLMPIYHQIAQTFDEGLGHHAVLIKADAGLGVERL 1739
>Contig190
          Length = 52,290
  Plus Strand HSPs:
 Score = 46 (21.3 bits), Expect = 19., P = 1.0
 Identities = 11/27 (40%), Positives = 19/27 (70%), Frame = +1
Query:   177 SREVTMSQDALLAALRLSAGSPGAALA 203
             S+ ++ S+ AL A++RLSA +  +A A
Sbjct: 48487 SKSLSNSRAALTASVRLSASTTASARA 48567
 Score = 45 (20.9 bits), Expect = 4.7, Poisson P(2) = 0.99
 Identities = 11/27 (40%), Positives = 13/27 (48%), Frame = +1
Query:   102 EHARLGGAKVVWVTDAALLTDAAANAL 128
             E ARL  A ++W     LL D   N L
Sbjct:  8032 EKARLALAMIIWQKPNLLLLDEPTNHL 8112
 Score = 44 (20.4 bits), Expect = 1.0, Poisson P(3) = 0.63
 Identities = 10/32 (31%), Positives = 15/32 (46%), Frame = +2
```

```
Query:    45 SRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDY 76
             +RY     P+G ++     R CQ +  G   DY
Sbjct:  2855 TRYYPLLHPRGGRAFARPRNCQNLP*GGDADY 2950
 Score = 40 (18.5 bits), Expect = 0.033, Poisson P(7) = 0.033
 Identities = 9/16 (56%), Positives = 11/16 (68%), Frame = +2
Query:   145 TREPERLLATLRSRCR 160
             TR+P RL A+L S  R
Sbjct: 15881 TRKPRRLRASLNSEHR 15928
 Score = 40 (18.5 bits), Expect = 0.033, Poisson P(7) = 0.033
 Identities = 9/25 (36%), Positives = 15/25 (60%), Frame = +1
Query:   167 PPEQYAVTWLSREVTMSQDALLAAL 191
             PP+    T  SR +T++    ++AAL
Sbjct: 40399 PPQTRVGTIFSRSLTVTGFTIMAAL 40473
 Score = 40 (18.5 bits), Expect = 0.033, Poisson P(7) = 0.033
 Identities = 10/36 (27%), Positives = 14/36 (38%), Frame = +2
Query:    51 QQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKGKN 86
             + P    +S G    GC    +  T PD         G+N
Sbjct:  3446 RSPADRRSEGKTVGCASRRHQTRPDSERQCQRPGRN 3553
 Score = 40 (18.5 bits), Expect = 0.033, Poisson P(7) = 0.033
 Identities = 15/41 (36%), Positives = 20/41 (48%), Frame = +1
Query:   220 LAYSVPSGDWYSLLAALNHEQAPARLHWLATLLMDALKRHH 260
             LA  VPS      LLA ++  Q  ARL     T +    LK+ +
Sbjct: 51634 LARRVPSAFKAKLLADMSDLQKSARLGQPDTTVAQWLKQRN 51756
 Score = 39 (18.1 bits), Expect = 0.51, Poisson P(7) = 0.40
 Identities = 12/45 (26%), Positives = 19/45 (42%), Frame = +3
Query:   104 ARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWFFLATREP 148
             A    GA ++    +   + + A  N L +     PP E W    +R P
Sbjct: 25044 APASGAGILTGMEVRVFSRAPNNKLSRLG*FPPLERWMSGLSRTP 25178
>Contig199
          Length = 81,564
  Minus Strand HSPs:
 Score = 51 (23.7 bits), Expect = 0.17, Poisson P(2) = 0.16
 Identities = 12/33 (36%), Positives = 18/33 (54%), Frame = -3
Query:   189 AALRLSAGSPGAALALFQGDNWQARETLCQALA 221
             AA+ LSAGS        + +G  W    + +C+A A
Sbjct: 13054 AAMILSAGSGSRITPVEKGITWFGLQPICRAAA 12956
 Score = 51 (23.7 bits), Expect = 0.17, Poisson P(2) = 0.16
 Identities = 14/62 (22%), Positives = 21/62 (33%), Frame = -2
Query:    53 PQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVV 112
             P+     C H R       +     H       L P  GKN  G    R+++     L               ++
Sbjct: 80819 PRSRFDCRHARPRYRHRRKQHTGRCRLCPRTGKNRCGRGLGRKLRRSLAPERDAPSKPLI 80640
Query:   113 WV 114
             W+
Sbjct: 80639 WM 80634
>Contig201
          Length = 92,813
  Minus Strand HSPs:
 Score = 54 (25.0 bits), Expect = 1.6, P = 0.80
 Identities = 13/29 (44%), Positives = 18/29 (62%), Frame = -2
Query:   188 LAALRLSAGSPGAALALFQGDNWQARETL 216
             L A++L+ G  GAA  LF  D  QA E++
Sbjct: 55888 LVAVKLNRGELGAAQLLFAPDETQALESV 55802
 Score = 47 (21.8 bits), Expect = 2.3, Poisson P(2) = 0.90
 Identities = 8/32 (25%), Positives = 17/32 (53%), Frame = -3
Query:    41 IYALSRYLLCQQPQGHKSCGHCRGCQLMQAGT 72
             ++ ++++   Q      KSC +CR  +  Q G+
Sbjct: 81990 VFGITKFCSRQTMYWRKSCAYCRNWKSSQHGS 81895
>Contig188
          Length = 44,251
  Plus Strand HSPs:
 Score = 45 (20.9 bits), Expect = 4.1, Poisson P(2) = 0.98
```

OU Neisseria Gonorrhoeae Sequence Blast Server Results

```
TBLASTN 1.3.9 [29-Oct-93]
Reference:  Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers,
and David J. Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol.
215:403-410.
Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.
Query=  ecoli.delta
        (343 letters)
Database:  /gono/abi/Gcphrap/auto_gono
           114 sequences; 2,133,469 total letters.
Searching.................................................done
                                                                     Smallest
                                                                      Poisson
                                                          Reading  High  Probability
Sequences producing High-scoring Segment Pairs:             Frame Score  P(N)      N
Contig188                                                      -2   233  1.2e-25   1
Contig183                                                      +2    46  0.26      4
Contig160                                                      +2    55  0.72      1
Contig163                                                      +2    51  0.77      2
Contig200                                                      +1    51  0.77      2
Contig149                                                      -1    44  0.85      3
Contig189                                                      +2    53  0.91      1
Contig126                                                      +2    47  0.91      2
Contig128                                                      +1    49  0.95      2
Contig129                                                      -3    45  0.98      2
Contig165                                                      +1    44  0.98      2
Contig190                                                      +1    51  0.99      1
Contig187                                                      -3    48  0.991     3
>Contig188
          Length = 44,251
  Minus Strand HSPs:
 Score = 233 (107.6 bits), Expect = 1.2e-25, P = 1.2e-25
 Identities = 56/186 (30%), Positives = 89/186 (47%), Frame = -2
Query:    10 RAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDWNAIFSLCQ 69
             R   +  L+   Y++ G + LL  E+ DA+R  A  QG+      ++ D + DWN +
Sbjct: 12126 RIDTDAPLKPLYVIHGEEELLRIEAVDALRAAAKKQGYLNREAYTADASFDWNELLQTAG 11947
Query:    70 AMSLFASRQTLLLLLPENGPNAAINEQLLTLTGLLHDDLLLIVRGNKLSKAQENAAWFTA 129
                LFA  + L L +P   P      E L        L +D + +V   KL K +   + WF A
Sbjct: 11946 NAGLFADLKLLELHIPNGKPGKNGGEALQDFAARLPEDTVTLVLLPKLEKTRLQSKWFAA 11767
Query:   130 LANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLS 189
             LA +         +     A LP+W+   R  ++ L ++  A  +      EGNLLA   Q +++L+
Sbjct: 11766 LAAKGEVWEAKPVGAAALPQWIRGRLDKIGLGIEADALALFAERVEGNLLAARQEIDKLA 11587
Query:   190 LLWPDG 195
             LL+P G
Sbjct: 11586 LLYPKG 11569
 Score = 73 (33.7 bits), Expect = 7.6e-08, Poisson P(2) = 7.6e-08
 Identities = 18/62 (29%), Positives = 28/62 (45%), Frame = -2
Query:   199 LPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGSEPVILLRTLQRELLLLV 258
             +    + AV + A F   F     A +       R    +L   L   EG EPV+LL   +   ++   L+
Sbjct: 11556 IDEAQTAVANVARFDAFQLAGAWMKADVPRVCRLLDGLEEEGEEPVLLLWAVAEDVRTLI 11377
Query:   259 NL 260
```

```
                L
Sbjct: 11376 RL 11371
 Score = 44 (20.3 bits), Expect = 0.90, Poisson P(3) = 0.60
 Identities = 11/55 (20%), Positives = 26/55 (47%), Frame = -2
Query:   277 RVWQNRRGMMGEALNRLSQTQLRQAVQLLTRTELTLKQDYGQSVWAELEGLSLLL 331
             R+W +++ +    A+ R+S  +L  A++    + +  +K        W   + L + L
Sbjct: 11319 RLWGDKQTLAPLAVKRISVVRLLDALKTCAQIDRIIKGAEDGDAWTVFKQLVVSL 11155
>Contig183
          Length = 34,103
  Plus Strand HSPs:
 Score = 46 (21.2 bits), Expect = 21., P = 1.0
 Identities = 9/29 (31%), Positives = 16/29 (55%), Frame = +2
Query:   293 LSQTQLRQAVQLLTRTELTLKQDYGQSVW 321
             L  T   +A +   RTE  +++ +G+ VW
Sbjct: 21089 LVSTSCNRAGKRACRTEREVRRQFGRDVW 21175
 Score = 43 (19.9 bits), Expect = 1.4, Poisson P(3) = 0.75
 Identities = 8/21 (38%), Positives = 12/21 (57%), Frame = +2
Query:   259 NLKRQSAHTPLRALFDKHRVW 279
             N  R+   +TP ++ F K R W
Sbjct: 10259 NAVRRFFNTPSKSCFSKARAW 10321
 Score = 43 (19.9 bits), Expect = 1.4, Poisson P(3) = 0.75
 Identities = 9/23 (39%), Positives = 14/23 (60%), Frame = +2
Query:    72 SLFASRQTLLLLLPENGPNAAIN 94
             +L A R      L+P++G N+ IN
Sbjct:  6584 NLSAGRVRTAFLMPKHGKNSKIN 6652
 Score = 43 (19.9 bits), Expect = 1.4, Poisson P(3) = 0.75
 Identities = 12/47 (25%), Positives = 19/47 (40%), Frame = +2
Query:   113 RGNKLSKAQENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLN 159
             RG+     +    A+W    +  R   +  C        +LP WVA   +    N
Sbjct: 10856 RGSYWLSSAVTASWRARMWARLRKGWCSHNANRRLPMWVAQPSSMEN 10996
 Score = 42 (19.4 bits), Expect = 0.30, Poisson P(4) = 0.26
 Identities = 10/34 (29%), Positives = 17/34 (50%), Frame = +2
Query:   154 RAKQLNLELDDAANQVLCYCYEGNLLALAQALER 187
             R K++   EL      +    CYC  +    L A+ +   E+
Sbjct:  6923 RYKEVIAELLAKGDAYYCYCSKEELEAMREKAEK 7024
>Contig160
          Length = 17,573
  Plus Strand HSPs:
 Score = 55 (25.4 bits), Expect = 1.3, P = 0.72
 Identities = 12/28 (42%), Positives = 16/28 (57%), Frame = +2
Query:   124 AAWFTALANRSVQVTCQTPEQAQLPRWV 151
             AA +  L +R   VT   P++AQ  RWV
Sbjct:  8054 AALYIRLCSRLPAVTAPIPQKAQKARWV 8137
 Score = 44 (20.3 bits), Expect = 3.8, Poisson P(2) = 0.98
 Identities = 11/32 (34%), Positives = 15/32 (46%), Frame = +1
Query:   163 DDAANQVLCYCYEGNLLALAQALERLSLLWPD 194
             DDA  +V       G + A    LE+  L +PD
Sbjct: 14512 DDAVKEVESLLMYGQIEAAMDVLEQAVLKYPD 14607
>Contig163
          Length = 24,139
  Plus Strand HSPs:
 Score = 51 (23.6 bits), Expect = 4.6, P = 0.99
 Identities = 9/21 (42%), Positives = 13/21 (61%), Frame = +2
Query:   134 SVQVTCQTPEQAQLPRWVAAR 154
             SV++ C +P  A LP W+  R
Sbjct: 13157 SVRLRCPSPSDATLPFWLRRR 13219
 Score = 46 (21.2 bits), Expect = 1.5, Poisson P(2) = 0.77
 Identities = 9/18 (50%), Positives = 12/18 (66%), Frame = +1
Query:   105 HDDLLLIVRGNKLSKAQE 122
             HDDLLL+++G      QE
Sbjct:  5695 HDDLLLVLKGAANKLVQE 5748
```

Following those BLAST hits is the sequence of the contig containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  delta prime
        (334 letters)

Database:  /usr/local/db/s_putrefaciens
           2430 sequences; 5,974,789 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                     Smallest
                                                                       Sum
                                                     Reading  High  Probability
Sequences producing High-scoring Segment Pairs:        Frame Score  P(N)      N

gsp_271                                                  +3    302  6.4e-27   1
gsp_387                                                  +1    192  1.9e-13   1


>gsp_271
        Length = 11,991

  Plus Strand HSPs:

 Score = 302 (106.3 bits), Expect = 6.4e-27, P = 6.4e-27
 Identities = 84/274 (30%), Positives = 132/274 (48%), Frame = +3

Query:     5 PWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG 64
             PWL     +  +    Q  +  HA L+       G + L    ++R  +C QP      CG C+
Sbjct:  1842 PWLDVPRQAFLTQLQTQKVPHAQLVGIDSAYGGELLSVFMARAAMCSQPTHTGGCGFCKS 2021

Query:    65 CQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAA 124
             CQL  AG HPD+Y +   E     + + VD +RE+  +L+  A+  G +V  +  +  L  A+
Sbjct:  2022 CQLFDAGNHPDFYQI--EADGHQIKVDQIRELCSRLSATAQQSGRRVAIIHHSERLNSAS 2195

Query:   125 ANALLKTLEEPPAETWFFLATREPERLLATLRSRC-RLHYLAGPPEQYAVTWLSREVTMS 183
             ANALLKTLEEP  +T    L +    P RL+AT+ SRC RL ++A P +       WL ++  +
Sbjct:  2196 ANALLKTLEEPGKDTLLLLHSDTPARLMATISSRCQRLPFVA-PSKTLIKNWLIQQCQIQ 2372

Query:   184 QDALLAALRLSAGSPGAALALFQGDNWQAR-ETLC---QALAYSVPSGDWYSLLAALNHE 239
             +D     L +  G    A +L         +R +TL     +  A S+ SG    + L  ++ +
Sbjct:  2373 EDVTWC-LSVVGGPLKLAESLQSNSTQPSRYQTLLGFRKDWAQSLSSGHLCASLLIISEQ 2549

Query:   240 QAPARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAEL 277
             Q     L  L  LL   L ++              +  L A++
Sbjct:  2550 QIIDALKVLYLLLRQILLKNGNQDAYVQAQIGNLAAKV 2663
```

```
>gsp_387
        Length = 3834

  Plus Strand HSPs:

 Score = 192 (67.6 bits), Expect = 1.9e-13, P = 1.9e-13
 Identities = 59/185 (31%), Positives = 86/185 (46%), Frame = +1

Query:    22 RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
             R HHA L     G+G  +L    ++ L C+       CG C  C  +  G   D   +
Sbjct:   562 RLHHAYLFTGTRGVGKTSLARLFAKGLNCETGVTASPCGVCGSCVEIAQGRFVDLIEV-- 735

Query:    82 EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWF 141
             +    T  VD  RE+ + +      G  KV  + +  +L+ ++ NALLKTLEEPP    F
Sbjct:   736 DAASRTK-VDDTRELLDNVQYRPTRGRFKVYLIDEVHMLSRSSFNALLKTLEEPPEHVKF 912

Query:   142 FLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMSQ-----DALLAALRLSAG 196
              LAT +P++L  T+ SRC    L    +Q   T L   +T  Q     +AL    + + G
Sbjct:   913 LLATTDPQKLPVTVLSRCLQFNLKSLTQQEIGTQLQHILTQEQLPFEHEALGLLAKSANG 1092

Query:   197 SPGAALAL 204
             S   AL+L
Sbjct:  1093 SMRDALSL 1116


Parameters:
  B=5

  ctxfactor=6.00
  E=10

  Query                          -----  As Used  -----    -----  Computed  ----
  Frame  MatID Matrix name      Lambda    K        H      Lambda    K        H
   +0      0   BLOSUM62          0.321   0.136   0.423    same     same    same
               Q=9,R=2           0.244   0.0300  0.180     n/a      n/a     n/a

  Query
  Frame  MatID  Length  Eff.Length     E     S W    T  X    E2     S2
   +0      0     334      334         10.   62 3   13 22  0.069   37
                                                      33  0.063   42


Statistics:

  Database:  /usr/local/db/s_putrefaciens
    Title:  /usr/local/db/s_putrefaciens
    Release date:  unknown
    Posted date:  10:07 AM EST Dec 15, 1998
    Format:  BLAST
  # of letters in database:  5,974,789
  # of sequences in database:  2430
  # of database sequences satisfying E:  2
  No. of states in DFA:  540 (57 KB)
  Total size of DFA:  97 KB (128 KB)
  Time to generate neighborhood:  0.00u 0.00s 0.00t  Elapsed: 00:00:00
  No. of threads or processors used:  1
  Search cpu time:  4.81u 0.01s 4.82t  Elapsed: 00:00:05
  Total cpu time:  4.84u 0.01s 4.85t  Elapsed: 00:00:05
  Start:  Wed Mar 17 09:14:58 1999   End:  Wed Mar 17 09:15:03 1999
```

Following those BLAST hits is the sequence of the <u>contig</u> containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  e coli delta
        (343 letters)

Database:  /usr/local/db/s_putrefaciens
           2430 sequences; 5,974,789 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                  Smallest
                                                                    Sum
                                                  Reading  High  Probability
Sequences producing High-scoring Segment Pairs:     Frame Score  P(N)       N

gsp_230                                                +2   564  1.1e-54    1
gsp_343                                                +1    70  0.999      1


>gsp_230
        Length = 21,837

  Plus Strand HSPs:

 Score = 564 (198.5 bits), Expect = 1.1e-54, P = 1.1e-54
 Identities = 135/343 (39%), Positives = 184/343 (53%), Frame = +2

Query:      2 IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
              +R+YP+QL    LN  L A YL+ G+DP LL+ S+D +RQ A  QGFEE        +    +W
Sbjct: 14210 MRVYPDQLSRHLNP-LHACYLIFGDDPWLLETSKDQIRQAAKRQGFEERVQLIQETGFNW 14386

Query:     62 NAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQLLTLTGLLHDDLLLIVRGNKLSKAQ 121
                 +      QAMSLF+SR+ + L LP     P A   +   L +L          D+LLI+ G KL+   Q
Sbjct: 14387 GDLTQEWQAMSLFSSRRIIELTLPSAKPGADGSAALQSLLQTPSPDVLLILEGPKLASEQ 14566

Query:    122 ENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLAL 181
               N+ WF   L +   + + C TPE  Q   RW+ +R       L L    A  +L     YEGNLLA
Sbjct: 14567 TNSKWFKTLDSLGIYLPCTTPEGDQFRRWLDSRIAHFKLNLQPDARAMLYSLYEGNLLAA 14746

Query:    182 AQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGS 241
               QA++ L LL P    +      +         D + FT F   DALL   +     A H+L QL   EG+
Sbjct: 14747 DQAMQLLQLLSPSKPIGADELSHYFEDQSRFTVFQLTDALLNNRQDSAQHMLAQLNGEGT 14926

Query:    242 EPVILLRTLQRELLLLVNLKRQSAH-TPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
                  ILL    L +EL  LL++LK  + A   +PL +LF  KHR+W   R+ +     AL  RLS   Q+
Sbjct: 14927 AMPILLWALFKELQLLLSLKSEQAQGSPLNSLFGKHRIWDKRKPLYQTALQRLSLAQIEH 15106
```

```
Query:   301 AVQLLTRTELTLKQDYGQSVWAELEGLSLLL---CHKPLADVFID 342
             +   ++ EL LKQ  G    W  L  L LL     H  LA + +D
Sbjct: 15107 MLAFASKLELNLKQ-LGHEDWTGLSHLCLLFDPKAHSHLAHINLD 15238


>gsp_343
        Length = 6977

  Plus Strand HSPs:

 Score = 70 (24.6 bits), Expect = 6.5, P = 1.00
 Identities = 33/127 (25%), Positives = 57/127 (44%), Frame = +1

Query:    19 AAYLLLGNDPLL--LQESQDAVRQVAAAQGFEEHH----TFSIDPNTDW-NAIFSLCQAM 71
             AA++L   N    +     E QDA +   ++ Q    +HH     TFSID N DW    + S     +
Sbjct:   466 AAHVLEDNGQQISGFIEVQDADKGQSSMQAMTDHHAAHGTFSIDVNGDWVYQLDSRRPDV 645

Query:    72 SLFASRQTLLLLLPENGPNAAINEQLLTLTGLLHDDLLLIVRGNKLSKAQENAAWFTALA 131
                 + +TLL   +   +   +    +E  +T+ G      ++        +L + Q  +A  T  A
Sbjct:   646 QALKAGETLLETITVHSADGTPHEVNITIHGQNDGAVISGADTGQLVEDQNVSAASTLEA 825

Query:   132 NRSVQVT 138
             +  + VT
Sbjct:   826 HGQLTVT 846


Parameters:
  B=5

  ctxfactor=6.00
  E=10

  Query                                 -----  As Used  -----     -----  Computed  ----
  Frame   MatID Matrix name             Lambda    K        H        Lambda    K        H
   +0       0   BLOSUM62                0.322   0.135    0.398     same     same     same
                Q=9,R=2                 0.244   0.0300   0.180      n/a      n/a      n/a

  Query
  Frame   MatID  Length  Eff.Length      E      S W    T  X     E2       S2
   +0       0      343       343         10.    62 3   13 22   0.067     37
                                                          33   0.063     42


Statistics:

  Database:  /usr/local/db/s_putrefaciens
    Title:  /usr/local/db/s_putrefaciens
    Release date:  unknown
    Posted date:  10:07 AM EST Dec 15, 1998
    Format:  BLAST
  # of letters in database:  5,974,789
  # of sequences in database:  2430
  # of database sequences satisfying E:  2
  No. of states in DFA:  531 (57 KB)
  Total size of DFA:  90 KB (128 KB)
  Time to generate neighborhood:  0.01u 0.00s 0.01t  Elapsed: 00:00:00
  No. of threads or processors used:  1
  Search cpu time:  4.46u 0.00s 4.46t  Elapsed: 00:00:04
  Total cpu time:  4.49u 0.00s 4.49t  Elapsed: 00:00:04
  Start:  Wed Mar 17 09:22:40 1999   End:  Wed Mar 17 09:22:44 1999
```

Following those BLAST hits is the sequence of the contig containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  e coli delta
        (343 letters)

Database:  /usr/local/db/v_cholerae
           694 sequences; 4,145,671 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                     Smallest
                                                                       Sum
                                                     Reading  High  Probability
Sequences producing High-scoring Segment Pairs:        Frame Score  P(N)      N

asm937                                                   +2   817  6.9e-82    1
asm843                                                   +3    68  0.9995     1


>asm937
       Length = 6994

  Plus Strand HSPs:

 Score = 817 (287.6 bits), Expect = 6.9e-82, P = 6.9e-82
 Identities = 168/338 (49%), Positives = 232/338 (68%), Frame = +2

Query:     2 IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
             +R+Y E+L   L++ L   YL+ GN+PLLLQE++ A+ + A AQGF E H FS D   DW
Sbjct:  1166 MRIYAEKLAESLHKTLYPIYLVFGNEPLLLQEAKTAIEKTAQAQGFLEKHRFSADAGLDW 1345

Query:    62 NAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQLLTLTGLLHDDLLLIVRGNKLSKAQ 121
             NA++  CQA+SLF+SRQ + + +PE+G NA    ++L  L G LH D+LL+V G KL+KAQ
Sbjct:  1346 NAVYDCCQALSLFSSRQLIEIEIPESGVNAQTAKELSALVGQLHQDILLLVIGPKLTKAQ 1525

Query:   122 ENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLAL 181
             ENAAWF  LA ++  V C TPE ++LP++V  R   L L+ D  A Q+L    +EGNL AL
Sbjct:  1526 ENAAWFKTLAQQACWVNCLTPELSRLPQFVQQRCFALGLKPDAEAVQMLAQWHEGNLFAL 1705

Query:   182 AQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGS 241
             AQ+LE+L+LL+PDG LTL R+E++++    HFTP+HW+DALL GK+ RA  IL+QL LE S
Sbjct:  1706 AQSLEKLALLYPDGLLTLVRLEESLSRHNHFTPYHWMDALLEGKANRAQRILRQLMLEES 1885

Query:   242 EPVILLRTLQRELLLLVNLKRQSAHTP-LRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
             EP+IL+RT Q+EL  L+  +++       L +LFD++RVWQNRR +   AL RL    L +
Sbjct:  1886 EPIILIRTAQKELTQLLKWQQERQQLGNLGSLFDRYRVWQNRRPLYSAALQRLPSRALLR 2065
```

```
Query:    301 AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADV 339
                V +LT+ EL   K   Y Q VW   L+ LSL   C+ P A++
Sbjct:   2066 LVGILTQAELLAKTQYEQPVWPILQQLSLECCN-PQANL 2179


>asm843
       Length = 26,802

  Plus Strand HSPs:

 Score = 68 (23.9 bits), Expect = 7.6, P = 1.00
 Identities = 22/63 (34%), Positives = 31/63 (49%), Frame = +3

Query:    115 NKLSKAQENAAWFT-ALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYC 173
              N LS   Q ++    T AL      +QV   Q PE A+   +WVA      ++   EL   +A   +
Sbjct: 15237 NVLSVYQPSSLVLTPALLMALIQVVKQAPELAKSLQWVAVGGARVAAELIHSARALGIPA 15416

Query:    174 YEG 176
              YEG
Sbjct: 15417 YEG 15425


Parameters:
  B=5

  ctxfactor=6.00
  E=10

  Query                                -----  As Used  -----     -----  Computed  ----
  Frame  MatID Matrix name            Lambda    K        H       Lambda    K        H
   +0      0   BLOSUM62               0.322   0.135   0.398      same    same     same
               Q=9,R=2                0.244   0.0300  0.180       n/a     n/a      n/a

  Query
  Frame  MatID  Length  Eff.Length     E    S W   T  X   E2     S2
   +0      0     343       343        10.   60 3  13 22  0.067   37
                                                     33  0.063   42


Statistics:

  Database:  /usr/local/db/v_cholerae
    Title:  /usr/local/db/v_cholerae
    Release date:  unknown
    Posted date:  12:58 PM EST Dec 11, 1998
    Format:  BLAST
  # of letters in database:  4,145,671
  # of sequences in database:  694
  # of database sequences satisfying E:  2
  No. of states in DFA:  531 (57 KB)
  Total size of DFA:  90 KB (128 KB)
  Time to generate neighborhood:  0.01u 0.00s 0.01t  Elapsed: 00:00:00
  No. of threads or processors used:  1
  Search cpu time:  3.25u 0.02s 3.27t  Elapsed: 00:00:03
  Total cpu time:  3.26u 0.03s 3.29t  Elapsed: 00:00:03
  Start:  Wed Mar 17 09:24:47 1999   End:  Wed Mar 17 09:24:50 1999
```

The top-scoring match came from this contig (up to 1000bp on either side of the hit are shown):

Following those BLAST hits is the sequence of the contig containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  delta prime
        (334 letters)

Database:  /usr/local/db/v_cholerae
           694 sequences; 4,145,671 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                 Smallest
                                                                   Sum
                                                  Reading  High  Probability
Sequences producing High-scoring Segment Pairs:     Frame Score  P(N)      N

asm894                                                 -1   394  8.1e-37   1
asm864                                                 -3   178  6.1e-12   1
asm959                                                 +3    79  0.37      1


>asm894
       Length = 19,711

  Minus Strand HSPs:

 Score = 394 (138.7 bits), Expect = 8.1e-37, P = 8.1e-37
 Identities = 106/313 (33%), Positives = 159/313 (50%), Frame = -1

Query:      4 YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
              YPWL P ++   A  AG+  A LIQA  G+G ++L+  ++R L+C   Q  + CG C
Sbjct:  18034 YPWLVPVWQPWQAGLAAGKISSATLIQASEGVGVESLVELMARTLMCTSSQS-EPCGFCH 17858

Query:     64 GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDA 123
               C LMQ+G HPD++ + PEK   ++ V+ +R++     E ++L G +++ +  A  + ++
Sbjct:  17857 SCGLMQSGNHPDFHVVKPEKIGKSITVEQIRQMNRIAQESSQLSGYRLIVIEPADAMNES 17678

Query:    124 AANALLKTLEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMS 183
              +ANALLKTLEEP     F L T   + LL T+ SRC+   L  P     V WL  +  ++
Sbjct:  17677 SANALLKTLEEPAPNCLFILVTSRIKHLLPTIVSRCQRLVLPAPTTALVVEWLKGQ-GIT 17501

Query:    184 QDALLAALRLSAGSPGAALA-LFQGDNWQARETLCQALAYSVPSGDWYSLLA--ALNHEQ 240
                A   AL L A SP   A + +G   +  E   Q +  +  SGD  + L     AL
Sbjct:  17500 TPAY--ALHLCADSPLKTRAFMLEGGAEKYHELESQLM--NALSGDVNAQLKCIALIDAD 17333

Query:    241 APARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQL 300
                  L+W+  +L DA K H G  Q         P  A LA   + S+L      +  + EQL
```

```
Sbjct: 17332 LTTHLYWVWCVLTDAQKIHFGVQQDY---YPPASAALAGRFTYSKLHVQTASLERLMEQL 17162

Query:   301 MSVTGINRELLITDLL 316
                +G+N ELL+    L
Sbjct: 17161 NQFSGLNTELLLLQWL 17114


>asm864
       Length = 23,778

  Minus Strand HSPs:

 Score = 178 (62.7 bits), Expect = 6.1e-12, P = 6.1e-12
 Identities = 46/143 (32%), Positives = 68/143 (47%), Frame = -3

Query:    22 RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
             R HHA L      G+G    +      ++ L C+          CG C   CQ +  G    D    +
Sbjct: 14509 RLHHAYLFSGTRGVGKTTIGRLFAKGLNCETGITATPCGQCATCQEIDQGRFVDLLEI-- 14336

Query:    82 EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWF 141
             +     T  V+   RE+ + +          G   KV   + +   +L+   + NALLKTLEEPP      F
Sbjct: 14335 DAASRTK-VEDTRELLDNVQYKPARGRFKVYLIDEVHMLSRHSFNALLKTLEEPPEYVKF 14159

Query:   142 FLATREPERLLATLRSRCRLHYL 164
              LAT +P++L   T+ SRC    +L
Sbjct: 14158 LLATTDPQKLPVTILSRCLQFHL 14090


>asm959
       Length = 15,780

  Plus Strand HSPs:

 Score = 79 (27.8 bits), Expect = 0.47, P = 0.37
 Identities = 35/115 (30%), Positives = 52/115 (45%), Frame = +3

Query:   174 TW-LSREVTMS---QDALLAALRLSAGSP---GAALALFQGDNWQARETLCQALAYSVPS 226
             +W LS  V+      QD L AA  L+   +     G  +AL  G      A +      ++A S P+
Sbjct:  1047 SWILSHRVSSELAHQDPLAAAFALAGATKDKCGTQMALVTG----ALKEDHVSVALSTPN 1214

Query:   227 GDWYSLLAALNHEQAPARLHWLATLLMDALKRH-HGAAQVTNVDVPGLVAELANHLSPSR 285
             G+W    +  +     A   +  W+ATL   D L R+  G +               V E+  HL  S
Sbjct:  1215 GEWGQTVKFVRRFSAQEQKEWIATLAADMLLRYLTGRSMFVGYSAVERVKEM--HLPSSV 1388

Query:   286 L 286
             L
Sbjct:  1389 L 1391
```

Parameters:

B=5

ctxfactor=6.00

E=10

Query Frame	MatID	Matrix name	----- As Used Lambda	K	H -----	----- Computed Lambda	K	H ----
+0	0	BLOSUM62	0.321	0.136	0.423	same	same	same
		Q=9,R=2	0.244	0.0300	0.180	n/a	n/a	n/a

Query

NCBI **BLAST Search Results** Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= ecoli.delta
(343 letters)

Searching...done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

```
                                                                    Score     E
Sequences producing significant alignments:                        (bits)  Value

gnl|PAGP|Paeruginosa_Contig52 Pseudomonas aeruginosa unfinished ...   139   7e-34
gnl|PAGP|Paeruginosa_Contig44 Pseudomonas aeruginosa unfinished ...    31   0.45
gnl|PAGP|Paeruginosa_Contig53 Pseudomonas aeruginosa unfinished ...    27   5.1
gnl|PAGP|Paeruginosa_Contig49 Pseudomonas aeruginosa unfinished ...    27   5.1
gnl|PAGP|Paeruginosa_Contig47 Pseudomonas aeruginosa unfinished ...    27   5.1

gnl|PAGP|Paeruginosa_Contig52 Pseudomonas aeruginosa unfinished fragment of complete ger
          Length = 872680

 Score =  139 bits (347), Expect = 7e-34
 Identities = 106/329 (32%), Positives = 155/329 (46%), Gaps = 8/329 (2%)
 Frame = -2

Query: 2      IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
              ++L P QL    L    L    Y++ G++PLL QE+ DA+RQ      + F E     F+ + N DW
Sbjct: 245226 MKLTPAQLAKHLQGPLAPVYVVSGDEPLLCQEACDAIRQACRERDFGERQVFNAEANFDW 245047

Query: 62     NAIFSLCQAMSLFASRQTLLLLLPENGPN---AAINEQXXXXXXXXXXXXXIVRGNKLS 118
               +      ++SLFA ++ + L LP    P      AAI ++                +     KL
Sbjct: 245046 GLLLEAGASLSLFAEKRLIELRLPSGKPGDKGAAILQEYLQRPPEDTVLLLGLP---KLD 244876

Query: 119    KAQENAAWFTAL--ANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEG 176
```

```
                 + +    W  AL    N +   +       +  QLP+W+  R  Q  L      A +++      EG
Sbjct: 244875 GSTQKTKWAKALIDGNAAQFIQVWPVDVHQLPQWIRQRLSQAGLSASPEALELIAARVEG 244696

Query: 177    NLLALAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQL 236
              NLLA AQ +E+L LL     ++      V+ AV D+A F  F  +DA L G++  AL IL+ L
Sbjct: 244695 NLLAAAQEIEKLKLLAEGNQIDAATVQAAVADSARFDVFGLIDAALGGEAAHALRILEGL 244516

Query: 237    RLEGSE-PVIXXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHR--VWQNRRGMMGEALNRL 293
              R EG E PVI                            PL   F + R  VW  RR ++  AL R
Sbjct: 244515 RGEGIEPPVILWGLAREIRLLAGLSQQYGQGIPLEKAFAQARPPVWDKRRPLLTRALQRH 244336

Query: 294    SQTQLRQAVQLLTRTELTLKQDYGQSVWAELEGLSLL 330
              S ++      Q+L   +L   Q  GQ+  +   GLSLL
Sbjct: 244335 SSSRWN---QMLRDAQLIDAQIKGQAPGSPWSGLSLL 244234
```

Score = 29.0 bits (63), Expect = 1.3
Identities = 20/50 (40%), Positives = 28/50 (56%)
Frame = +2

```
Query: 10     RAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNT 59
              RA+L +GL     LLL +      +Q S+ AVR++AA  G        T    DP+T
Sbjct: 87335  RARLAQGLSLTDLLLEH---AIQPSRSAVRRLAAGGGLRLDGTPVSDPDT 87475
```

gnl|PAGP|Paeruginosa_Contig44 Pseudomonas aeruginosa unfinished fragment of complete ger
Length = 203793

Score = 30.5 bits (67), Expect = 0.45
Identities = 19/54 (35%), Positives = 25/54 (46%)
Frame = +3

```
Query: 274    DKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTRTELTLKQDYGQSVWAELEGL 327
              D    + Q R    +G   L +L QTQ       V LL      ++        Y    V+A LEGL
Sbjct: 157899 DGEAIAQLRTDELGGLLRKLRQTQQMALVGLLRNQDVATSLGYLARVYARLEGL 158060
```

gnl|PAGP|Paeruginosa_Contig53 Pseudomonas aeruginosa unfinished fragment of complete ger
Length = 1300758

Score = 27.0 bits (58), Expect = 5.1
Identities = 18/53 (33%), Positives = 33/53 (61%), Gaps = 4/53 (7%)
Frame = +2

```
Query: 156    KQLNLELDDAANQVLC-YCYEGNLLALAQALERLSLLWPDGKL---TLPRVEQAVND 208
              K+ ++   + AA    LC + + GN+   LA   +ERL+++ P G +       LP+   + V+D
Sbjct: 462347 KRGSIRFNSAAIMSLCRHDWPGNVRELANLVERLAIMHPYGVIGVGELPKKFRHVDD 462517
```

gnl|PAGP|Paeruginosa_Contig49 Pseudomonas aeruginosa unfinished fragment of complete ger
(15-MAR-99)
Length = 476032

Score = 27.0 bits (58), Expect = 5.1
Identities = 14/37 (37%), Positives = 24/37 (64%), Gaps = 7/37 (18%)
Frame = -2

```
Query: 124  AAWFTALANRSVQVTCQTP-------EQAQLPRWVAARAKQLNL 160
            AA    A+AN  V +   +T              E+ +LPRW++ R   +++L
Sbjct: 2694 AALMLAMANMRVLLAARTKRPSLPAFEEVRLPRWLSGRTMKISL 2563
```

NCBI	BLAST Search Results	Entrez	?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= deltaprime.ecoli
(334 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

Sequences producing significant alignments:	Score (bits)	E Value
gnl|PAGP|Paeruginosa_Contig50 Pseudomonas aeruginosa unfinished ...	115	9e-27
gnl|PAGP|Paeruginosa_Contig53 Pseudomonas aeruginosa unfinished ...	62	1e-10
gnl|PAGP|Paeruginosa_Contig47 Pseudomonas aeruginosa unfinished ...	29	1.00
gnl|PAGP|Paeruginosa_Contig45 Pseudomonas aeruginosa unfinished ...	29	1.00
gnl|PAGP|Paeruginosa_Contig46 Pseudomonas aeruginosa unfinished ...	29	1.3
gnl|PAGP|Paeruginosa_Contig52 Pseudomonas aeruginosa unfinished ...	28	2.9

```
 gnl|PAGP|Paeruginosa_Contig50 Pseudomonas aeruginosa unfinished fragment of complete ger
          Length = 798876

 Score =  115 bits (286), Expect = 9e-27
 Identities = 84/323 (26%), Positives = 139/323 (43%), Gaps = 11/323 (3%)
 Frame = +2

Query: 4      YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
              YPW +  + +L     Q      HA L+     G+G  AL      +  LLCQ+P    +CG C+
Sbjct: 521618 YPWQQALWSQLGGRAQHA---HAYLLYGPAGIGKRALAEHWAAQLLCQRPAAAGACGECK 521788

Query: 64     GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
              CQL+ AGTHPDY+ L PE+ +   + VD VR++    + + A+LGG KVV +
Sbjct: 521789 ACQLLAAGTHPDYFVLEPEEAEKPIRVDQVRDLVGFVVQTAQLGGRKVVLLEPAEAMNVN 521968
```

```
Query: 124     XXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMS 183
                        EEP  +T   L + +P RLL T++SRC       P    ++ WL+R +
Sbjct: 521969  AANALLKSLEEPSGDTVLLLISHQPSRLLPTIKSRCVQQACPLPGAAASLEWLARALPDE 522148

Query: 184     QDXXXXXXXXXXXXXXXXXXXFQGDNWQ---------ARETLCQALAYSVPSGDWYSLLA 234
                                      G   +          ++ L Q +A S  +  W
Sbjct: 522149  PAEALEELLALSGGSPLTAQRLHGQGVREQRAQVVEGVKKLLKQQIAASPLAESW----- 522313

Query: 235     ALNHEQAPARLHWLATLLMDALKRH--HGAAQVTNVDVPGLVAELANHLSPSRLQAILGD 292
                  N    P    W     +  L+    H    +   D+  ++  L +    +++ A+
Sbjct: 522314  --NSVPLPLLFDWFCDWTLGILRYQLTHDEEGLGLADMRKVIQYLGDKSGQAKVLAMQDW 522487

Query: 293     VCHIREQLMSVTGINRELLITDLLLRIEHYLQPG 326
                +   R+++++   +NR LL+  LL++      PG
Sbjct: 522488  LLQQRQKVLNKANLNRVLLLEALLVQWASLPGPG 522589
```

Score = 30.1 bits (66), Expect = 0.58
Identities = 17/36 (47%), Positives = 22/36 (60%)
Frame = +2

```
Query: 13      KLVASYQAGRGHHALLIQALPGMGDDALIYALSRYL 48
               +L  +    RGH  LLI+ LPGMG   L +AL+R L
Sbjct: 613469  RLALACLLARGH--LLIEDLPGMGKTTLSHALARVL 613570
```

Score = 28.2 bits (61), Expect = 2.2
Identities = 18/69 (26%), Positives = 28/69 (40%)
Frame = +1

```
Query: 14      LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
               L+A+  AG   H    + L  +GD         + +  QP  H   G CRG + +
Sbjct: 670210  LLAALLAGYLAHLFCRRRLSLVGDMYRAMRAREFHMVYQPIIHLDTGECRGVEALVRWQR 670389

Query: 74      PDYYTLAPE 82
               PD   + P+
Sbjct: 670390  PDRSQVRPD 670416
```

Score = 26.2 bits (56), Expect = 8.6
Identities = 22/72 (30%), Positives = 32/72 (43%)
Frame = +2

```
Query: 258     RHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQLMSVTGINRELLITDLLL 317
               RHHG      +    LV  L +HL P   ++  G V H  E+      ++R L    L+
Sbjct: 795185  RHHGEEATVGMAHGALVDVLGHHLHPDLHRSAPG-VVHRGEEGHQFADMDR-LAEDHLIH 795358

Query: 318     RIEHYLQPGVVL 329
               R  H++  GV L
Sbjct: 795359  RQGHHVASGVAL 795394
```

gnl|PAGP|Paeruginosa_Contig53 Pseudomonas aeruginosa unfinished fragment of complete ger
Length = 1300758

Score = 62.1 bits (148), Expect = 1e-10
Identities = 69/268 (25%), Positives = 103/268 (37%), Gaps = 12/268 (4%)
Frame = +2

```
Query: 14      LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
               L+ +    R HHA L     G+G   +   L++ L C+       CG C  C+ +  G
```

NCBI **BLAST Search Results** Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= ecoli.delta
(343 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

```
                                                                    Score    E
Sequences producing significant alignments:                         (bits)  Value

gnl|Sanger|S.typhiContig1564 Salmonella typhi unfinished fragmen...   563   e-161
gnl|Sanger|S.typhiContig1088 Salmonella typhi unfinished fragmen...    28   2.0
gnl|Sanger|S.typhiContig1954.0 Salmonella typhi unfinished fragm...    28   2.0
gnl|Sanger|S.typhiContig2054 Salmonella typhi unfinished fragmen...    26   6.0

 gnl|Sanger|S.typhiContig1564 Salmonella typhi unfinished fragment of complete genome
           Length = 3596

 Score =  563 bits (1435), Expect = e-161
 Identities = 279/343 (81%), Positives = 298/343 (86%)
 Frame = +3

Query: 1    MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 60
            MIRLYPEQLRAQLNE LRAAYLLLGNDPLLLQESQDA+R   AA+QGFEEHH F++DP+TD
Sbjct: 1500 MIRLYPEQLRAQLNEWLRAAYLLLGNDPLLLQESQDAIRLAAASQGFEEHHAFTLDPSTD 1679

Query: 61   WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKA 120
            W ++FSLCQAMSLFASRQTL+L LPENGPNAA+NEQ              IVRGNKL+KA
Sbjct: 1680 WGSLFSLCQAMSLFASRQTLVLQLPENGPNAAMNEQLATLSELLHDDLLLIVRGNKLTKA 1859

Query: 121  QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA 180
            QENAAW+TALA+RSVQV+CQTPEQAQLPRWVAARAK  NL+LDDAANQ+LCYCYEGNLLA
```

```
Sbjct: 1860 QENAAWYTALADRSVQVSCQTPEQAQLPRWVAARAKAQNLQLDDAANQLLCYCYEGNLLA 2039

Query: 181  LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 240
            LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG
Sbjct: 2040 LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 2219

Query: 241  SEPVIXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
            SEPVI                 QSAHTPLRALFDKHRVWQNRR M+G+AL RL    QLRQ
Sbjct: 2220 SEPVILLRTLQRELLLLVNLKRQSAHTPLRALFDKHRVWQNRRPMIGDALQRLHPAQLRQ 2399

Query: 301  AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFIDG 343
            AVQLLTRTE+TLKQDYGQSVWA+LEGLSLLLCHK LADVFIDG
Sbjct: 2400 AVQLLTRTEITLKQDYGQSVWADLEGLSLLLCHKALADVFIDG 2528


 gnl|Sanger|S.typhiContig1088 Salmonella typhi unfinished fragment of complete genome
           Length = 2112

 Score = 27.8 bits (60), Expect = 2.0
 Identities = 14/38 (36%), Positives = 21/38 (54%)
 Frame = -1

Query: 270 RALFDKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTR 307
           R LF +HR   + RRG  G+     + Q +LR   +  +TR
Sbjct: 963 RKLFQRHRPLRQRRGRRGKDHQLIFQPRLRDNLCTVTR 850


 gnl|Sanger|S.typhiContig1954.0 Salmonella typhi unfinished fragment of complete genome
           Length = 3497

 Score = 27.8 bits (60), Expect = 2.0
 Identities = 14/36 (38%), Positives = 23/36 (63%)
 Frame = +3

Query: 54   SIDPNTDWNAIFSLCQAMSLFASRQTLLLLLPENGP 89
            +++P T W+    S  QAMS FA +++   +LLP + P
Sbjct: 1464 TVNPVTPWSP*ISRYQAMSAFARQKS--VLLPSSSP 1565


 gnl|Sanger|S.typhiContig2054 Salmonella typhi unfinished fragment of complete genome
           Length = 6017

 Score = 26.2 bits (56), Expect = 6.0
 Identities = 18/47 (38%), Positives = 28/47 (59%), Gaps = 12/47 (25%)
 Frame = -1

Query: 263  QSAHTPLRALFDKHRV------WQNRRG------MMGEALNRLSQTQLRQAVQLLTRTE 309
             +S ++T LRAL+DKH V            NR G          M   A + + + ++    V   + +TE
Sbjct: 5450 RSIYTDLRALYDKHNVAGITASQTNREGGASEVATMMHAADNIEKVRIADLVITINKTE 5274


CPU time:     0.05 user secs.      0.01 sys. secs         0.06 total secs.

  Database: Unfinished Salmonella typhi
    Posted date:  Dec 15, 1998 12:07 PM
  Number of letters in database: 4,464,430
  Number of sequences in database:  1746

Lambda     K      H
   0.321    0.134    0.00
```

NCBI # BLAST Search Results **Entrez** **?**

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ.
Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= deltaprime.ecoli
(334 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

```
                                                                   Score     E
Sequences producing significant alignments:                        (bits)  Value

gnl|Sanger|Y.pesits_Contig51 Yersinia pestis unfinished fragment...  284   9e-78
gnl|Sanger|Y.pesits_Contig774 Yersinia pestis unfinished fragmen...   63   6e-11
gnl|Sanger|Y.pesits_Contig695 Yersinia pestis unfinished fragmen...   28   1.8
gnl|Sanger|Y.pesits_Contig675 Yersinia pestis unfinished fragmen...   28   2.3
gnl|Sanger|Y.pesits_Contig777 Yersinia pestis unfinished fragmen...   27   3.0
gnl|Sanger|Y.pesits_Contig701 Yersinia pestis unfinished fragmen...   26   6.8

 gnl|Sanger|Y.pesits_Contig51 Yersinia pestis unfinished fragment of complete genome
            Length = 20197

 Score =  284 bits (720), Expect = 9e-78
 Identities = 147/334 (44%), Positives = 192/334 (57%), Gaps = 6/334 (1%)
 Frame = -1

Query: 1     MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG 60
              M WYPWL    + +LV   +   GRGHHALL+ +LPG G+DALIYALSR+L+CQQ QG KSCG
Sbjct: 15274 MNWYPWLNAPYRQLVGQHSTGRGHHALLLHSLPGNGEDALIYALSRWLMCQQRQGEKSCG 15095

Query: 61    HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXX 120
               C  C+LM AG HPD+Y L PEKGK+++GV+ VR++ +KL  HA+ GGAKVVW+
Sbjct: 15094 ECHSCRLMLAGNHPDWYVLTPEKGKSSIGVELVRQLIDKLYSHAQQGGAKVVWLPHAEVL 14915
```

```
Query: 121   XXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLS--- 177
                         EEPP +T+F L   +P  LLATLRSRC   YLA P     + WL+
Sbjct: 14914 TDAAANALLKTLEEPPEKTYFLLDCHQPASLLATLRSRCFYWYLACPDTAICLQWLNLQW 14735

Query: 178   --REVTMSQDXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAA 235
               R++ +                       Q + W  R  LC  L ++    D  SLL
Sbjct: 14734 RKRQIPVEPVAMLAALKLSEGAPLAAERLLQPERWSIRSALCSGLREALNRSDLLSLLPQ 14555

Query: 236   LNHEQAPARLHWLATLLMDALKRHHGAAQ-VTNVDVPGLVAELANHLSPSRLQAILGDVC 294
             LNH+ A  RL WL++LL+DALK   GA +   N D   LV +LA+  +   L  +    +
Sbjct: 14554 LNHDDAAERLQWLSSLLLDALKWQQGAGEFAVNQDQLPLVQQLAHIAATPVLLQLAKQLA 14375

Query: 295   HIREQLMSVTGINRELLITDLLLRIEHYLQPGVVLPVPHL 334
             H R QL+SV G+NRELL+T+ LL  E  L  G    +P L
Sbjct: 14374 HCRHQLLSVVGVNRELLLTEQLLSWETALSTGTYSTLPSL 14255
```

gnl|Sanger|Y.pesits_Contig774 Yersinia pestis unfinished fragment of complete genome
Length = 66020

Score = 62.8 bits (150), Expect = 6e-11
Identities = 40/144 (27%), Positives = 60/144 (40%)
Frame = +2

```
Query: 21   GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
            GR HHA L     G+G  ++   L++ L C+        CG C  CQ ++ G   D   +
Sbjct: 2714 GRIHHAYLFSGTRGVGKTSIARLLAKGLNCETGITATPCGTCANCQEIEQGRFVDLIEI- 2890

Query: 81   PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETW 140
                 +   V+  RE+ + +      G  KV  +                  EEPPA
Sbjct: 2891 --DAASRTKVEDTRELLDNVQYAPARGRFKVYLIDEVHMLSRHSFNALLKTLEEPPAHVK 3064

Query: 141  FFLATREPERLLATLRSRCRLHYL 164
            F LAT +P++L  T+ SRC   +L
Sbjct: 3065 FLLATTDPQKLPVTILSRCLQFHL 3136
```

gnl|Sanger|Y.pesits_Contig695 Yersinia pestis unfinished fragment of complete genome
Length = 43655

Score = 28.2 bits (61), Expect = 1.8
Identities = 8/13 (61%), Positives = 11/13 (84%)
Frame = +3

```
Query: 54   QGHKSCGHCRGCQ 66
            +GH +CGHCR C+
Sbjct: 9102 EGHITCGHCRNCR 9140
```

gnl|Sanger|Y.pesits_Contig675 Yersinia pestis unfinished fragment of complete genome
Length = 1090

Score = 27.8 bits (60), Expect = 2.3
Identities = 15/41 (36%), Positives = 21/41 (50%)
Frame = -2

```
Query: 213 RETLCQALAYSVPSGDWYSLLAALNHEQAPARLHWLATLLM 253
           +E+ C +  Y   S   YS+L+A  H   P RL W   +LM
Sbjct: 786 QESECLSCYYQDQSYLHYSILSACLHHWIPDRLRWPEYMLM 664
```

NCBI **BLAST Search Results** Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

```
Query= ecoli.delta
        (343 letters)

Searching.................................................done
```

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits)  Value

gnl|Sanger|Y.pesits_Contig803 Yersinia pestis unfinished fragmen...   447   e-127
gnl|Sanger|Y.pesits_Contig689 Yersinia pestis unfinished fragmen...    27   3.1
gnl|Sanger|Y.pesits_Contig701 Yersinia pestis unfinished fragmen...    27   3.1
gnl|Sanger|Y.pesits_Contig798 Yersinia pestis unfinished fragmen...    27   5.3
gnl|Sanger|Y.pesits_Contig795.0 Yersinia pestis unfinished fragm...    26   6.9
gnl|Sanger|Y.pesits_Contig765 Yersinia pestis unfinished fragmen...    26   6.9

 gnl|Sanger|Y.pesits_Contig803 Yersinia pestis unfinished fragment of complete genome
            Length = 177561

 Score =  447 bits (1138), Expect = e-127
 Identities = 223/342 (65%), Positives = 263/342 (76%)
 Frame = +1

Query: 1     MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 60
            MIR+YPEQL AQL+EGLRA YLL GN+PLLLQESQD +R+VA+    F EH +F++D +T+
Sbjct: 50068 MIRIYPEQLVAQLHEGLRACYLLCGNEPLLLQESQDHIRRVASQHDFTEHFSFALDAHTE 50247

Query: 61    WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKA 120
            W  IFSLCQA+SLFASRQTLLL  P++G  A I+EQ              I+R NKL+KA
Sbjct: 50248 WEHIFSLCQALSLFASRQTLLLSFPDSGLTAPISEQLVKLSGLLHPDILLILRANKLTKA 50427
```

```
Query: 121   QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA 180
             QEN+AWF AL+   V V+CQTPEQAQLPRWV+ARAK LNL +DDAA Q+LCYCYEGNLLA
Sbjct: 50428 QENSAWFKALSKNGVFVSCQTPEQAQLPRWVSARAKSLNLNVDDAAIQLLCYCYEGNLLA 50607

Query: 181   LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 240
             L+QALERLSLL+PDGKLTLP+VEQAVNDAAHFTP+HW+DALLMGKSKRA HILQQL+ E
Sbjct: 50608 LSQALERLSLLYPDGKLTLPKVEQAVNDAAHFTPYHWLDALLMGKSKRAWHILQQLQQED 50787

Query: 241   SEPVIXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
             SEPVI                 Q     PLRALFD+H++WQNRR MM +AL RLS  QL+Q
Sbjct: 50788 SEPVILLRTVQRELLLLLALKRQMEQVPLRALFDQHKIWQNRRPMMTQALQRLSLQQLQQ 50967

Query: 301   AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFID 342
             AV LLT+ E+ LKQDYGQS+W ELE LS+L+C K L + F D
Sbjct: 50968 AVHLLTQMEIRLKQDYGQSIWPELETLSMLMCGKTLPESFFD 51093
```

gnl|Sanger|Y.pesits_Contig689 Yersinia pestis unfinished fragment of complete genome
Length = 32290

Score = 27.4 bits (59), Expect = 3.1
Identities = 23/72 (31%), Positives = 32/72 (43%), Gaps = 11/72 (15%)
Frame = -2

```
Query: 267   TPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTRTELTLKQDY---------- 316
             T   L   + V Q    + G  L+RLS   LRQ V L+ +   + L    +
Sbjct: 20130 TLANXLMGYYPVQQGEIXLDGRPLSRLSHQVLRQGVALVQQDPVVLADSFFXNITXGRDL 19951

Query: 317   -GQSVWAELEGLSLLLCHKPLAD 338
              Q VW  LE + L     + L D
Sbjct: 19950 SEQQVWEALETVQLAPLVRTLPD 19882
```

gnl|Sanger|Y.pesits_Contig701 Yersinia pestis unfinished fragment of complete genome
Length = 67923

Score = 27.4 bits (59), Expect = 3.1
Identities = 19/58 (32%), Positives = 27/58 (45%)
Frame = -3

```
Query: 142   PEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLWPDGKLTL 199
             PEQA+           A    +LEL    A  ++      GNLL +AQA      + WP   K+ +
Sbjct: 36676 PEQAETVVLAEGYATAQSLELLLPAAVIIAAIDAGNLLPVAQAFR---IYWPAAKIII 36512
```

gnl|Sanger|Y.pesits_Contig798 Yersinia pestis unfinished fragment of complete genome
Length = 112126

Score = 26.6 bits (57), Expect = 5.3
Identities = 13/33 (39%), Positives = 19/33 (57%), Gaps = 1/33 (3%)
Frame = +1

```
Query: 28    PLLLQESQDAVRQVA-AAQGFEEHHTFSIDPNTD 60
             PL++        +VR +A    +GF EH     I+PN D
Sbjct: 94696 PLVIHNFLQSVRLLADGMRGFNEHCALGIEPNRD 94797
```

gnl|Sanger|Y.pesits_Contig795.0 Yersinia pestis unfinished fragment of complete genome
Length = 71341

Score = 26.2 bits (56), Expect = 6.9

4420–4449 Nucleic Acids Research, 1996, Vol. 24, No. 22

Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*

Ralf Himmelreich, Helmut Hilbert+, Helga Plagens, Elsbeth Pirkl, Bi-Chen Li§ and Richard Herrmann*

Zentrum für Molekulare Biologie Heidelberg, Mikrobiologie, Universität Heidelberg, 69120 Heidelberg, Germany

Received August 22, 1996; Revised and Accepted October 10, 1996 DDBJ/EMBL/GenBank accession no. U00089

ABSTRACT

The entire genome of the bacterium *Mycoplasma pneumoniae* M129 has been sequenced. It has a size of 816 394 base pairs with an average G+C content of 40.0 mol%. We predict 677 open reading frames (ORFs) and 39 genes coding for various RNA species. Of the predicted ORFs, 75.9% showed significant similarity to genes/proteins of other organisms while only 9.9% did not reveal any significant similarity to gene sequences in databases. This permitted us tentatively to assign a functional classification to a large number of ORFs and to deduce the biochemical and physiological properties of this bacterium. The reduction of the genome size of *M.pneumoniae* during its reductive evolution from ancestral bacteria can be explained by the loss of complete anabolic (e.g. no amino acid synthesis) and metabolic pathways. Therefore, *M.pneumoniae* depends in nature on an obligate parasitic lifestyle which requires the provision of exogenous essential metabolites. All the major classes of cellular processes and metabolic pathways are briefly described. For a number of activities/functions present in *M.pneumoniae* according to experimental evidence, the corresponding genes could not be identified by similarity search. For instance we failed to identify genes/proteins involved in motility, chemotaxis and management of oxidative stress.

INTRODUCTION

The bacterium *Mycoplasma pneumoniae* has a genome size of ~800 kb and completely lacks a cell wall. The bacterium is surrounded by a cytoplasmic membrane only, which contains cholesterol as an indispensable component. *Mycoplasma pneumoniae* is a human pathogen, causing 'atypical pneumonia' (1) usually in older children and young adults. As a surface parasite, it attaches to the host's respiratory epithelium by means of a differentiated terminal structure termed attachment organelle or tip structure. For a long time, research activities mainly focused on pathogenicity-related topics such as studies on cytadherence (2), vaccination and diagnosis (3). *Mycoplasma pneumoniae* was not considered as an organism suitable for basic studies partly because of its fastidious growth requirements and partly because of the lack of established standard genetic tools like conjugation or transformation with self-replicating vectors (4). These disadvantages can be compensated now to a large extent by the methods of molecular biology.

Morowitz pointed out in 1984, that mycoplasmas would be suitable candidates for defining the genetic constitution of a minimal self-replicating cell (5). The advantage of these bacteria for such studies (6,7), mainly due to their small genome size, was so obvious that several initiatives were started to sequence five different mycoplasma genomes: *Mycoplasma genitalium* (8,9), *M.pneumoniae* (10), *Mycoplasma capricolum* (11), *Mycoplasma mycoides* (12) and a species from the related genus *Ureaplasma*, *Ureaplasma urealyticum* (13). So far, only the complete sequence of the *M.genitalium* genome has been published (9) which, with 580 070 bp, is the smallest bacterial genome known so far. In the genus *Mycoplasma*, *M.pneumoniae* and *M.genitalium* are the closest related species. We report in this publication the complete nucleotide sequence of the genome of *M.pneumoniae*, which thus provides information on a second small bacterial genome. All *M.pneumoniae* genes which had been already sequenced were reanalyzed except for the P1 operon (14). Our sequencing strategy, early results and a detailed description of *M.pneumoniae* as an experimental system have been recently published (10).

MATERIALS AND METHODS

Mycoplasma strain

The strain *Mycoplasma pneumoniae* M129 (ATTC 29342) in the 18th broth passage was used to construct an ordered cosmid library containing the complete genome (15). This cosmid library was the basis for the DNA sequence analysis. We selected this specific bacterial strain because it has been used in cytadherence and pathogenicity studies (2,16,17). The strain in the 20th broth passage was still infectious in hamsters (H. Brunner, unpublished data).

DNA sequencing

Using the enzymatic dideoxy chain-termination method (18), the sequence data for this study were exclusively generated on a fluorescent-based sequence-gel reader (Model 373A, Applied Biosystems). Sequencing strategies and methods were as described in Hilbert *et al.* (10).

*To whom correspondence should be addressed. Tel: +49 6221 54 68 27; Fax: +49 6221 54 58 93; Email: r.herrmann@mail.zmbh.uni-heidelberg.de

Present addresses: +QIAGEN GmbH, 40724 Hilden, Germany and §Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

Computer assisted analysis

Sequence assembly, map drawing and multiple alignments were done with the *Lasergene* program package (DNA STAR).

Other analyses were performed with the *HUSAR* (Heidelberg Unix Sequence Analysis Resources) program package release 4.0 at the German Cancer Research Center, Heidelberg, Germany. This package is based on the *GCG* program package version Unix-8.1 of the Genetics Computer Group, Wisconsin. For searching the DNA and protein databases [*SWISS-PROT* (19) and *PIR* (20)] the *FASTA* (21) and *BLAST* (22) programs (*BLASTX*, *BLASTN* and *BLASTP*) were used. Conserved motifs in proteins and peptides were identified by using the program *PROSITE* (23). Open reading frames (ORFs) were calculated by the program *FRAMES* allowing AUG (or GUG, UUG) as start codons using the Mycoplasma translation table where UGA codes for tryptophan (24). The G+C content was calculated by the program *WINDOW*. Codon usage was performed with the program *CODONFREQUENCY*.

The programs *TopPred 1.1.1* (Manuel G. Carlos, Ecole Normale Superieure, Laboratoire de Genetique Moleculaire, Paris, France) and PSORT (25) (http://psort.nibb.ac.jp/) were used for the prediction of transmembrane domains and the membrane topology of proteins.

Each ORF analysis is accessible as a *File Maker Pro* (Claris) database which can be accessed at our world wide web (www) site (http://zmbh.uni-heidelberg.de/M_pneumoniae). It contains, besides genome and cosmid position of each ORF/gene, data about expression, availibility of antibodies, comments, literature, prosite patterns, amino acid composition and database search homology scores. All the annotations in this paper were done on the basis of the highest score values.

Accession number

The complete *M.pneumoniae* sequence has been annotated in GenBank (NCBI) with the accession number U00089.

RESULTS AND DISCUSSION

The strategy and methodology for sequencing the complete genome has been described by us recently (10). A total of 2 415 202 nucleotides primary sequence data were provided by 6385 sequencing reactions. Each strand of the genome was completely sequenced at least once. The direct sequencing approach, combining primer walking with a limited shotgun strategy based on a complete cosmid and plasmid library considerably facilitated the assembly of the individual sequences to the entire genome sequence. The average redundancy of the sequencing was 2.95 (calculated for both strands). This very low redundancy was achieved by the use of 5095 oligonucleotides.

The complete *M.pneumoniae* genome has a size of 816 394 bp and a G+C content of 40.0 mol%. Altogether 677 open reading frames (ORFs) and 39 genes coding for various RNA species were predicted. All ORFs were sorted into categories according to their proposed functions (Tables 1 and 2; Fig. 1). Only 333 ORFs (49.2%) were functionally assigned, based on significant sequence similarities to genes or proteins from other organisms with known functions (e.g. ribosomal proteins) or at least known categories of function (e.g. proteins involved in cytadherence). Significant similarities to proteins without known function from other bacteria, mostly *M.genitalium*, were shown for 181 proposed ORFs (26.7%). We also included in this group those *M.pneumoniae* proteins which were identified in protein extracts of *M.pneumoniae* by monospecific antibodies or by the N-terminal amino acid sequences of enriched proteins (26,27). The group of ORFs without significant similarity or without indication for their *in vivo* expression comprised 109 members (16.1%); 42 of them carry characteristic motifs, which are not sufficient for defining a function. Examples of such motifs are the leucine zipper (29 cases; refered to all predicted ORFs), the typical prokaryotic lipoprotein sequence pattern (46 cases) or ATP- and GTP-binding sites (73 cases). In addition all predicted gene products were analyzed by programs for structure predictions, e.g. coiled/coiled structures (29 cases) or transmembrane segments (275 cases). The latter result supports the analysis of cell fractionation experiments which indicate that the membrane fraction contains ~50% of the total proteins estimated by SDS-PAGE. About 8% of the genome is composed of repetitive DNA elements RepMP1, RepMP2/3, RepMP4 and RepMP5, while only 67 of all predicted ORFs (9.9%) code for a product without any similarity to a known RNA or protein.

Finally, 58 gene families were defined comprising 298 proteins with at least two but frequently with more paralogs; these are proteins with similarities within the same species (see www pages).

The proposed ORFs are not equally distributed over the genome. A lower coding density coincides with regions of lower or higher G+C content than the average. There are regions with a G+C content of up to 56 mol%. These regions code almost exclusively for the gene P1 and gene ORF6 of the P1 operon, the repetitive DNA sequences RepMP4, RepMP2/3, RepMP5 and tRNAs (for details see www pages).

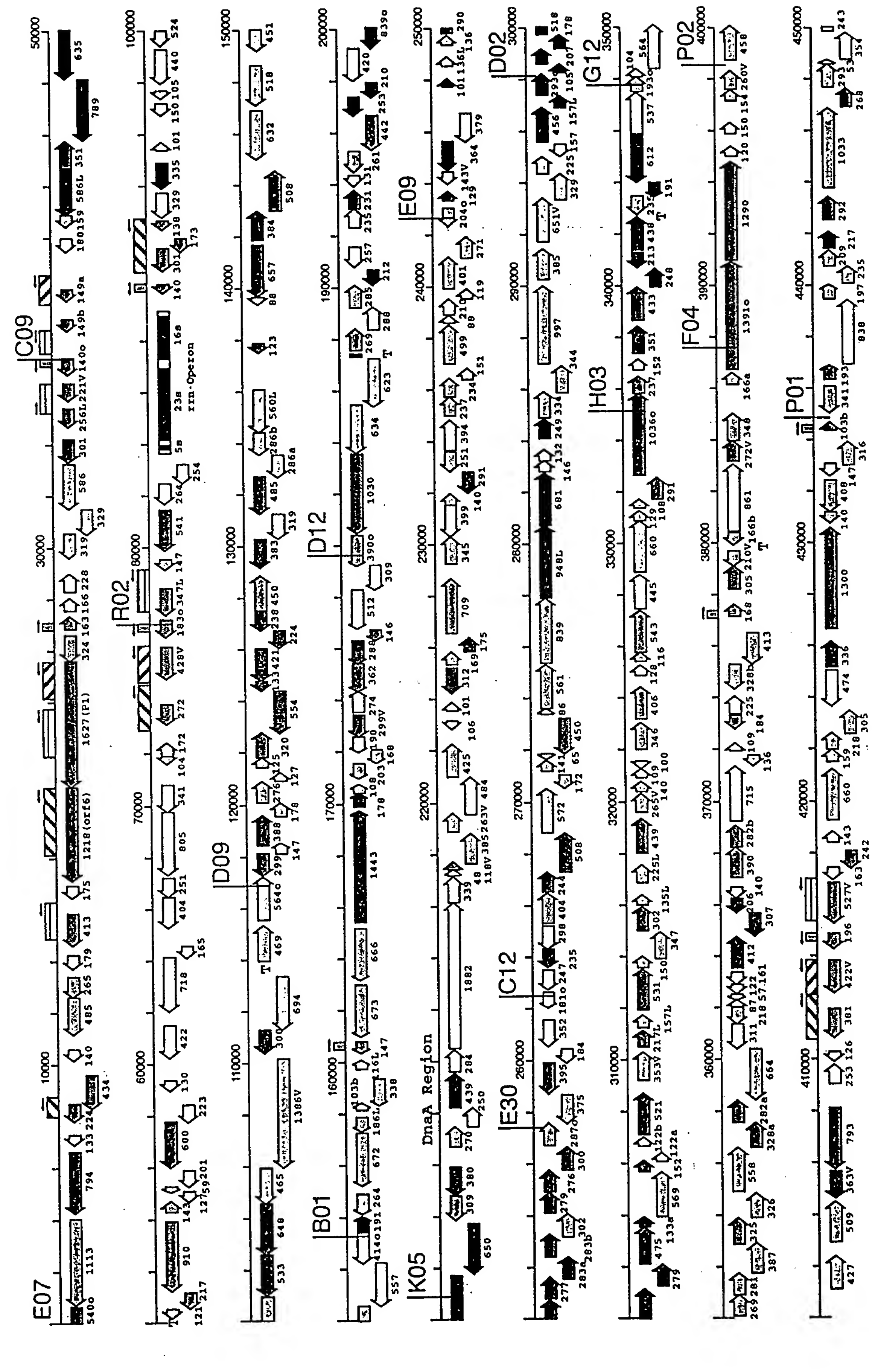
The P1 protein, the main adhesin, is essential for adherence of *M.pneumoniae* to its host cell (28) and the ORF6 gene product which is only found as a cleavage product, namely a 40 and 90 kDa protein, instead of the expected 130 kDa protein, is involved in an as yet unknown manner in cytadherence (14). Gene P1 contains a copy each of RepMP2/3 and RepMP4 and gene ORF6 one of RepMP5 (29). In addition, several copies of each of these repetitive DNA sequences can easily be recognized by their relative high G+C content (Fig. 2).

At the other extreme is the proposed origin of replication around nucleotide position 205 000 (pcosMPK05, dnaA region), with a G+C content of only 26 mol% (10).

Other regions with a low G+C content do not show a similar obvious coding pattern, but proposed ORFs coding for lipoproteins or the hsd modification/restriction system are frequently located in these regions.

The total length of all coding regions is 724 174 bp. The average coding density of 88.7% was calculated for the *M.pneumoniae* genome which gives an average gene size of 1011 bp. Similar

Figure 1. (Following two pages) The gene map of the complete *M.pneumoniae* genome. The arrows indicate the position and the size of the predicted ORFs. The colour refers to the functional category in which the ORFs are sorted. The complete name of an ORF can be deduced by the cosmid name above the horizontal scale-line and the number below the arrows (e.g. the ORF name of the first complete arrow in this figure is E07_orf1113). Rectangles above the scale-line indicate the size and the position of different repetitive DNA sequences (see also Table 4).




Biosynthesis of cofactors; prosthetic groups
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acids and phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory funktion
Replication
Translation
Transcription
Transport / binding proteins
Other categories / Hypothetical
REP derived ORFs
Unknown
REPMP1
REPMP2/3
REPMP4
REPMP5
T: tRNAs
rRNA and other RNA

Table 1. Predicted functions and classification of all *M.pneumoniae* ORFs

• Biosynthesis of cofactors, prosthetic groups and carrier - Folic acid [5]		
F10_orf160	[a]MG228	dihydrofolate reductase (dhfr); LACLA
H10_orf506	MG213	dihyrofolate reductase (dyr) homolog protein; ENTFC
D12_orf269	MG013	5,10-methylene-tetrahydrofolate dehydrogenase (mtd1); HAEIN
D02_orf406	MG394	serine hydroxymethyltransferase (glyA); ACTAC
H91_orf164	MG245	5-formyl tetrahydrofolate cyclo-ligase (HI0858) homolog; HAEIN
• Biosynthesis of cofactors, prosthetic groups and carrier - Heme and porphyrin [1]		
H91_orf453	MG259	possible protoporphyrinogen oxidase (hemK); ECOLI
• Biosynthesis of cofactors, prosthetic groups and carrier - Thioredoxin [2]		
A65_orf102	MG124	thioredoxin (trx); YEAST
K04_orf315	MG102	thioredoxin reductase (trxB); EUBAC
• Cell envelope - Membranes, lipoproteins and porines [42]		
A05_orf1244	MG307	putative lipoprotein, MG307 homolog, MYCGE
A05_orf252	MG440	putative lipoprotein, MG440 homolog, MYCGE
A65_orf251a	MG440	putative lipoprotein, MG440 homolog, MYCGE
A65_orf787o	MG260	putative lipoprotein, MG260 homolog, MYCGE
A65_orf794	MG260 (MG185)	putative lipoprotein, MG260 homolog, MYCGE
D02_orf217L	MG395 (MG068)	putative lipoprotein, MG395 homolog, MYCGE
D02_orf302	MG068 (MG395)	putative lipoprotein, MG068 homolog, MYCGE
D02_orf439	MG068 (MG395)	putative lipoprotein, MG068 homolog, MYCGE
D02_orf521	MG395 (MG068)	putative lipoprotein, MG395 homolog, MYCGE
D02_orf531	MG395 (MG068)	putative lipoprotein, MG395 homolog, MYCGE
D09_orf123	-	putative lipoprotein
D09_orf485	MG045	putative lipoprotein, MG045 homolog, MYCGE
D09_orf657	MG040	putative lipoprotein, MG040 homolog, MYCGE
D12_orf231	-	putative lipoprotein
E07_orf301	MG186	putative lipoprotein, MG186 homolog, MYCGE
E07_orf794	MG260 (MG185)	putative lipoprotein, MG260 homolog, MYCGE
E09_orf101	marginal MG440	putative lipoprotein
E09_orf129	-	putative lipoprotein
E09_orf276	MG440	putative lipoprotein, MG440 homolog, MYCGE
E09_orf277	MG440	putative lipoprotein, MG440 homolog, MYCGE
E09_orf279	MG439	putative lipoprotein, MG439 homolog, MYCGE
E09_orf283a	MG439	putative lipoprotein, MG439 homolog, MYCGE
E09_orf283b	MG439	putative lipoprotein, MG439 homolog, MYCGE
E09_orf290	MG439	putative lipoprotein, MG439 homolog, MYCGE
E09_orf300	MG439	putative lipoprotein, MG439 homolog, MYCGE
F11_orf760	MG260 (MG185)	putative lipoprotein, MG260 homolog, MYCGE
G07_orf454	MG095	putative lipoprotein, MG095 homolog, MYCGE
G12_orf305	MG348	putative lipoprotein, MG348 homolog, MYCGE
GT9_orf760	MG185	putative lipoprotein, MG185 homolog, MYCGE
GT9_orf798	MG260	putative lipoprotein, MG260 homolog, MYCGE
H08_orf1005	MG321	putative lipoprotein, MG321 homolog, MYCGE
H08_orf1325	MG309	putative lipoprotein, MG309 homolog, MYCGE
H08_orf150	MG307	putative lipoprotein, MG307 homolog, MYCGE
H08_orf237	MG307	putative lipoprotein, MG307 homolog, MYCGE
H91_orf102	MG260	putative lipoprotein, MG260 homolog, MYCGE
H91_orf253	-	putative lipoprotein
P01_orf101	-	putative lipoprotein
P02_orf1300	MG338	putative lipoprotein, MG338 homolog, MYCGE
P02_orf793	MG260	putative lipoprotein, MG260 homolog, MYCGE
R02_orf533	MG067	putative lipoprotein, MG067 homolog, MYCGE
R02_orf541	MG260	putative lipoprotein, MG260 homolog, MYCGE
VXpSPT7_orf320	MG149	putative lipoprotein, MG149 homolog, MYCGE
• Cell envelope - Surface structures and cytadherence [8]		
E07_orf1627	MG191 (MG192)	adhesin P1 (orf5, P1 operon); MYCPN
E07_orf1218	MG192 (MG191)	hypothetical 130K protein (orf6; P1 operon); MYCPN
H08_orf274	MG318	30K adhesin-related protein; MYCPN
H08_orf1018	MG312	cytadherence accessory protein (hmw1); MYCPN
F10_orf1818	MG218	cytadherence accessory protein (hmw2); MYCPN
H08_orf672	MG317	cytadherence accessory protein (hmw3); MYCPN
D02_orf1036o	MG386	protein P200; MYCPN
F10_orf405	MG217	protein P65; MYCPN
• Cell envelope - Surfaces polysaccharides, lipopolysaccharides and antigens [4]		
A65_orf399V	MG137	YefE protein homolog; ECOLI
B01_orf299V	MG025	TrsB protein; YEREN
D09_orf299	MG060	hypothetical protein YWDF homolog; BACSU
G12_orf282b	MG356	LicA protein homolog; HAEIN
• Cellular processes - Cell division [2]		
F10_orf380	MG224	cell division protein (ftsZ); BACSU
K05_orf709	MG457	cell division protein (ftsH); BACSU
• Cellular processes - Cell killing [1]		
VXpSPT7_orf424	MG146	hemolysin (hlyC) homolog protein; HAEIN
• Cellular processes - Chaperones [7]		
A05_orf595	MG305	heat shock protein DnaK, ERYRH
C09_orf217	MG201	heat shock protein GrpE, HAEIN

Table 1. *Continued*

D02_orf116	MG393	heat shock protein GroES; BACSU
D02_orf543	MG392	heat shock protein GroEL; BACSU
D12_orf390o	MG019	heat shock protein DnaJ; BACSU
C09_orf910	MG200	DnaJ homolog protein, MYCCA
K05_orf309	MG002	DnaJ homolog protein; YEAST
• **Cellular processes - Detoxification [1]**		
D12_orf442	MG008	possible thiophene and furan oxidation protein (tdhF); BACSU
• **Cellular processes - Protein and peptide secretion [9]**		
A05_orf348	MG297	cell division protein (ftsY); ECOLI
D09_orf450	MG048	signal recognition particle protein (ffh); MYCMY-
G07_orf808	MG072	preprotein translocase (secA); BACSU
GT9_orf477	MG170	preprotein translocase secY subunit; MYCCA
A65_orf581	MG138	GTP-binding membrane protein (lepA); HAEIN
F10_orf444	MG238	trigger factor (tig); HAEIN
H10_orf184	MG210	prolipoprotein signal peptidase (lsp); STACA
G07_orf389b	MG086	prolipoprotein diacylglyceryl transferase (lgt); ECOLI
F11_orf339	MG270	lipoate protein ligase (lplA); ECOLI
• **Central intermediary metabolism - Other [5]**		
A05_orf241a	MG293	glycerophosphoryl diester phosphodiesterase (glpQ); BACSU
A05_orf320	MG299	phosphotransacetylase (pta); BACSU
D09_orf508	MG038	glycerol kinase (glpK), HAEIN
G12_orf390	MG357	acetate kinase (ackA); BACSU
H03_orf237	MG385	glycerophosphoryl diester phosphodiesterase (glpQ); STAAU
• **Central intermediary metabolism - Phosphorous compounds [1]**		
G12_orf184	MG351	inorganic pyrophosphatase (ppa); THEAC
• **Energy metabolism - Aerobic [3]**		
K05_orf312	MG460	L-lactate dehydrogenase (ldh); MYCHY
D09_orf384	MG039	aerobic glycerol-3-phospate dehydrogenase (glpD); ECOLI
F11_orf479	MG275	NADH oxidase (nox); ENTFA
• **Energy metabolism - Amino acids and amines [5]**		
F10_orf309	-	carbamate kinase (EC 2.7.2.2) (arcC); PSEAE
H03_orf438	-	arginine deiminase (arcA); PSEPU
H10_orf198	-	arginine deiminase (arcA); MYCCA
H10_orf238	-	arginine deiminase (arcA); MYCCA
H10_orf273o	-	ornithine carbamoyl transferase (otc1); ECOLI
• **Energy metabolism - Anaerobic [1]**		
H03_orf351	-	NADP-dependent alcohol dehydrogenase (adh); THEBR
• **Energy metabolism - ATP-proton motive force interconversion [9]**		
C12_orf293o	MG405	ATP synthase A chain (atpB); MYCGA
D02_orf207	MG403	ATP synthase B chain (atpF); MYCGA
D02_orf105	MG404	ATP synthase C chain (atpE); MYCGA
C12_orf157L	MG406	ATP synthase protein I (atpI); MYCGA
D02_orf518	MG401	ATP synthase alpha chain (atpA); MYCGA
D02_orf475	MG399	ATP synthase beta chain (atpD); MYCGA
D02_orf279	MG400	ATP synthase gamma chain (atpG); MYCGA
D02_orf178	MG402	ATP synthase delta chain (atpH); MYCGA
D02_orf133a	MG398	ATP synthase epsilon chain (atpC); MYCGA
• **Energy metabolism - Glycolysis [10]**		
A05_orf337	MG301	glycerladehyde-3-phosphate dehydrogenase(gap), CLOPA
A05_orf409	MG300	phosphoglycerate kinase (pgk); THEMA
B01_orf288	MG023	fructose-bisphosphate aldolase (tsr); BACSU
C12_orf244	MG431	triosephosphate isomerase (tim); ECOLI
C12_orf456	MG407	enolase (eno) (EC 4.2.1.11); PLAFA
C12_orf508	MG430	phosphoglycerate mutase (pgm); BACSU
H10_orf328	MG215	6-phosphofructokinase (pfk); ECOLI
H10_orf508	MG216	pyruvate kinase (pyk); LACLA
K04_orf430	MG111	phosphoglucose isomerase B (pgiB); BACST
R02_orf300	MG063	1-phosphofructokinase (fruK); HAEIN
• **Energy metabolism - Pentose Phosphate pathway [2]**		
P02_orf242	-	L-ribulose-5-phosphate 4-epimerase (araD); ECOLI
R02_orf648	MG066	transketolase 1 (TK 1; tktB); RHOSH
• **Energy metabolism - Pyruvate DHase [4]**		
F11_orf327	MG273	pyruvate dehydrogenase E1-beta subunit (pdhB); ACHLA
F11_orf358a	MG274	pyruvate dehydrogenase E1-alpha subunit (pdhA); ACHLA
F11_orf402	MG272	dihydrolipoamide acetyltransferase component (E2) (pdhC); ACHLA
F11_orf457	MG271	dihydrolipoamide dehydrogenase (pdhD); BACST
• **Energy metabolism - Sugars [5]**		
D02_orf152	MG396	galactose-6-phosphate isomerase subunit (LacA); STRMU
D09_orf224	MG050	deoxyribose-phosphate aldolase (deoC); MYCPN
D09_orf554	MG053	phosphomannomutase (cpsG); MYCPI
E09_orf364	-	mannitol-1-phosphate 5-dehyrogenase (EC 1.1.1.17)(mtlD); STRMU
K04_orf215L	MG112	D-ribulose-5-phosphate 3 epimerase (cfxE); ALCEU

Table 1. *Continued*

• **Fatty acid and phospholipid metabolism [9]**		
A65_orf227	MG114	phosphatidylglycerophosphate synthase (pgsA); HAEIN
C09_orf600	-	carnitine palmitoyltransferase II precursor(cpt2); HUMAN
E30_orf395	MG437	CDP-diglyceride synthetase (cdsA); HAEIN
F11_orf84	MG287	(acyl carrier protein; STRGA)
G12_orf272V	MG344	triacylglycerol lipase (lip) 3; MYCMY
G12_orf328a	MG368	fatty acid/phospholipid synthesis protein (plsX); ECOLI
H08_orf289	MG310	triacylglycerol lipase (lip) 3; Mycoplasma sp
H10_orf266	MG212	1-acyl-sn-glycerol-3-phosphate acyltransferase (plsB); YEAST
P01_orf268	MG327	triacylglycerol lipase (lip) 2; MYCMY
• **Purines, pyrimidines, nucleosides and nucleotides - 2'-Deoyribonucleotide metabolism [3]**		
F10_orf328	MG227	thymidylate synthase (thyA); STAAU
F10_orf339	MG229	ribonucleotide reductase 2 (nrdF); SALTY
F10_orf721	MG231	ribonucleoside-diphosphate reductase (nrdE); SALTY
• **Purines, pyrimidines, nucleosides and nucleotides - Nucleotide and nucleoside interconversions [2]**		
C12_orf235	MG434	uridylate kinase (pyrH); ECOLI
H03_orf213	MG382	uridine kinase (udk); HAEIN
• **Purines, pyrimidines, nucleosides and nucleotides - Purine ribonucleotide biosynthesis [3]**		
D09_orf388	MG058	phosphoribosylpyrophosphate synthetase (prs); SYNP
GT9_orf215	MG171	adenylate kinase (adk); BACST
K04_orf239	MG107	5'guanylate kinase (gmk); HAEIN
• **Purines, pyrimidines, nucleosides and nucleotides - Salvage of nucleosides and nucleotides [9]**		
B01_orf178	MG030	uracil phosphoribosyltransferase (upp); STRSL
B01_orf191	MG034	thymidine kinase (tdk); BACSU
D09_orf133	MG052	cytidine deaminase (cdd); MYCPI
D09_orf238	MG049	purine-nucleoside phosphorylase (deoD); ECOLI
D09_orf421	MG051	thymidine phosphorylase (deoA); MYCPI
F11_orf133	MG276	adenine phosphoribosyltransferase (apt); HAEIN
K05_orf175	MG458	hypoxanthine-guanine phosphoribosyltransferase (HPT); LACLA
P01_orf217	MG330	cytidylate kinase (cmk); BACSU
D12_orf210	MG006	thymidylate kinase (CDC8) homolog, MYCGE
• **Purines, pyrimidines, nucleosides and nucleotides - Sugar-nucleotide biosynthesis and conversions [2]**		
A65_orf335	MG118	UDP-glucose 4-epimerase (galE); STRTR
K05_orf291	MG453	UDP-glucose pyrophosphorylase (gtaB); BACSU
• **Pyridine nucleotide synthesis [1]**		
H03_orf248	MG383	probable NH(3)-dependent NAD(+) synthetase (outB); BACSU
• **Regulatory function [8]**		
B01_orf362	MG024	hypothetical protein (yyaF) homolog; BACSU
C09_orf351	MG205	protein hrcA homolog, BACSU
D02_orf291	MG387	GTP-binding protein era homolog; STRMU
F11_orf733	MG278 (MG376)	stringent response protein SpoT; ECOLI
H03_orf433	MG384	GTP-binding protein (obg); BACSU
K04_orf726	MG104	virulence associated protein homolog (vacB); HAEIN
P01_orf193	MG335	hypothetical protein YihA (era like) homolog; ECOLI
P01_orf292	MG329	hypothetical protein HI0136 (era like) homolog; HAEIN
• **Replication - DNA replication, restriction, modification, recombination and repair [46]**		
A65_orf711	MG122	DNA topoisomerase I (topA); BACSU
A19_orf291	MG262	DNA polymerase I (polI, 5'-3' exonuclease) homolog; STRPN
A19_orf872	MG261	DNA polymerase III alpha subunit (dnaE); HAEIN
B01_orf1443	MG031	DNA polymerase III (dnaE) alpha chain (3'-5' exonuclease); BACSU
K05_orf380	MG001	DNA polymerase III beta subunit (dnaN); STAAU
D12_orf253	MG007	DNA polymerase III subunit delta' (holB); ECOLI
C12_orf681	MG420(C-Term:MG419)	DNA polymerase III subunit gamma and tau (dnaX); ECOLI
G07_orf473	MG094	replicative DNA helicase (dnaC); BACSU
H91_orf620	MG250	DNA primase (dnaG); BACSU
D12_orf212	MG010	DNA primase motif (dnaG); CLOAB
H91_orf658	MG254	DNA ligase (lig); ECOLI
G07_orf166	MG091	single-stranded DNA binding protein (ssb); HAEIN
K05_orf439	MG469	chromosomal replication initiator protein (dnaA); MYCCA
P02_orf336	MG339	recombination protein (recA); STAAU
C09_orf635	MG203	topoisomerase IV subunit B (parE), BACSU
C09_orf789	MG204	topoisomerase IV subunit A (parC), BACSU
K05_orf650	MG003	DNA gyrase subunit B (gyrB); MYCPN
K05_orf839o	MG004	DNA gyrase subunit A (gyrA); STAAU
G12_orf206	MG358	Holliday junction DNA helicase (ruvA); ECOLI
G12_orf307	MG359	Holliday junction DNA helicase (ruvB); HAEIN
H91_orf715	MG244	DNA helicase II (mutB1); HAEIN
H91_orf529	MG244	DNA helicase pcrA homolog; STAAU
F10_orf286	MG235	endonuclease IV (nfo); ECOLI
C12_orf948L	MG421	excinuclease ABC subunit A (uvrA); ECOLI
G07_orf657	MG073	excinuclease ABC subunit B (uvrB); ECOLI
C09_orf586L	MG206	excinuclease ABC subunit C (uvrC), BACSU
G12_orf412	MG360	UV protection protein (mucB); ECOLI
A19_orf277	MG(M2)	formamidopyrimidine-DNA glycosylase (fpg); BACFI
A65_orf306	-	PrrB homolog protein, ECOLI

Table 1. *Continued*

D09_orf383	MG047	S-adenosylmethionine synthetase 2 (metX); ECOLI
G07_orf240	MG097	uracil DNA glycosylase (ung); ECOLI
C12_orf249	-	restriction-modification enzyme subunit S1B (hsdS); MYCPU
GT9_orf238	-	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
GT9_orf319V	MG184	adenine-specific methyltransferase EcoRI (mte1); ECOLI
H03_orf191	MG380	glucose inhibited division protein (gidB); ECOLI
H03_orf612	MG379	glucose inhibited division protein (gidA); ECOLI
H10_orf145L	-	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
H10_orf187V	-	HsdS1B protein homolog; MYCPU
H91_orf206	-	Type 1 restriction enzyme (hsdR) homolog; ECOLI
H91_orf268	-	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
H91_orf330	-	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
H91_orf376	-	Type 1 restriction enzyme (hsdR) homolog; ECOLI
H91_orf543	-	type I restriction enzyme (hsdM); ECOLI
P02_orf363V	-	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
R02_orf335	-	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
E30_orf375	MG438	MG438 homolog, MYCGE
• **Transcription - Degradation of RNA [2]**		
G12_orf282a	MG367	ribonuclease III (rnc); ECOLI
K05_orf118V	MG465	RNaseP C5 chain (rnpA); MYCCA
• **Transcription - RNA synthesis, modification and DNA transcription [11]**		
GT9_orf327	MG177	RNA polymerase alpha core subunit (rpoA); BACSU
G12_orf1391o	MG341	RNA polymerase beta subunit (rpoB); BACSU
F04_orf1290	MG340	DNA-directed RNA polymerase beta' chain (rpoC); THEMA
B01_orf146	MG022	DNA-directed RNA polymerase delta subunit (rpoE); BACSU
H91_orf499	MG249	RNA polymerase sigma-A factor (sigA); BACSU
F11_orf160	MG282	transcription elongation factor (greA); RICPR
D09_orf320	MG054	transcription antitermination factor (nusG); BACSU
E07_orf540o	MG141	N-utilization substance protein A homolog (nusA); BACSU
C12_orf450	MG425	ATP-dependent RNA helicase (deaD); HAEIN
H08_orf409	MG308	ATP-dependent RNA helicase (deaD); ECOLI
D12_orf1030	MG018	hypothetical helicase Yb95 homolog; YEAST
• **Translation - Amino acyl tRNA synthetases and tRNA modification [24]**		
A05_orf900	MG292	alanyl-tRNA synthetase (alaS); ECOLI
H03_orf537	MG378	arginyl-tRNA synthetase (argS); BRELA
K04_orf455o	MG113	asparaginyl-tRNA synthetase (asnS); ECOLI
D09_orf557	MG036	aspartyl-tRNA synthetase (aspS); THEAQ
H91_orf437	MG253	cysteinyl-tRNA synthetase (cysS); BACSU
K05_orf484	MG462	glutamyl-tRNA synthetase (gltX); BACST
H91_orf449	MG251	glycyl-tRNA synthetase (grs1); YEAST
B01_orf414o	MG035	histidyl-tRNA synthetase (hisS); STREQ
G12_orf861	MG345	isoleucine-tRNA ligase (ileS); STAAU
F11_orf793o	MG266	leucyl-tRNA synthetase (leuS); BACSU
A65_orf489	MG136	lysyl-tRNA synthetase (lysS); BACSU
G12_orf311	MG365	methionyl-tRNA formyltransferase (fmt); ECOLI
B01_orf512	MG021	methionyl-tRNA synthetase (metS); BACST
G07_orf188	MG083	peptidyl-tRNA hydrolase homolog (pth); HAEIN
C09_orf341	MG194	phenylalanyl-tRNA synthetase alpha-subunit (pheS); BACSU
C09_orf805	MG195	phenylalanyl-tRNA synthetase beta chain (pheT); BACSU
GT9_orf243V	MG182	pseudouridylate synthase I (hisT); ECOLI
F11_orf483	MG283	putative prolyl-tRNA synthetase (YHI0; proS); YEAST
D12_orf420	MG005	seryl-tRNA synthetase (serS); BACSU
G12_orf564	MG375	threonyl-tRNA synthetase (thrSv); BACSU
K05_orf210	MG445	tRNA (guanine-N1)-methyltransferase (trmD); HUMAN
A65_orf346	MG126	tryptophanyl-tRNA synthetase (trpS); HAEIN
K05_orf399	MG455	tyrosyl tRNA synthetase (tyrS); BACCA
P01_orf838	MG334	valyl-tRNA synthetase (valS); BACST
• **Translation - Degradation of proteins, peptides and glycopeptides [8]**		
B01_orf309	MG020	proline iminopeptidase (pip); NEIGO
D02_orf445	MG391	nonspecified aminopeptidase; MYCSA
D09_orf319	MG046	o-sialoglycoprotein endopeptidase (gcp); PASHA
F10_orf795	MG239	ATP-dependent protease (lon); BACSU
G12_orf715	MG355	ATP-dependent protease binding subunit (clpB) homolog; HAEIN
GT9_orf611	MG183	oligoendopeptidase F (pepF); LACLA
H03_orf193o	MG377	MG377 homolog (put. zinc protease), MYCGE
P01_orf354	MG324	X-Pro dipeptidase (pepX); LACDE
• **Translation - Protein modification and translation factors [15]**		
GT9_orf78	MG173	initiation factor 1 (infA); BACSU
VXpSPT7_orf617	MG142	protein synthesis initiation factor 2 (infB); BACST
C09_orf201	MG196	translation initiation factor IF3 (infC); MYCFE
G07_orf688	MG089	elongation factor G (fus); THEAQ
B01_orf190	MG026	elongation factor P (efp) homolog; HAEIN
C12_orf298	MG433	elongation factor Ts (tsf); SPICI
K05_orf394	MG451	elongation factor TU (tuf); MYCGE
H91_orf359V	MG258	peptide chain release factor 1 (RF1; prfA);BACSU
E30_orf184	MG435	ribosome releasing factor (frr); HAEIN
GT9_orf248	MG172	methionine amino peptidase (map); BACSU
K04_orf216	MG106	polypeptide deformylase (def); HAEIN
K04_orf259	MG108	protein phoshatase 2C homolog; YEAST

Table 1. *Continued*

K04_orf389	MG109	probable protein serine/threonine kinase; CAEEL
K05_orf151	MG448	pilB homolog (fragment); HAEIN
C12_orf157	MG408	peptide methionine sulfoxide reductase (pmsR), ECOLI
• **Translation - Ribosomal proteins: synthesis and modification [53]**		
G07_orf226	MG082	ribosomal protein L1 (rpL1); BACST
VXpSPT7_orf287a	MG154	ribosomal protein L2 (rpL2); MYCCA
VXpSPT7_orf287b	MG151	ribosomal protein L3 (rpL3); MYCCA
VXpSPT7_orf212	MG152	ribosomal protein L4 (rpL4); MYCCA
GT9_orf180b	MG163	ribosomal protein L5 (rpL5); HAEIN
GT9_orf184	MG166	ribosomal protein L6 (rpL6); MYCCA
G12_orf122	MG362	ribosomal protein L7/L12 ('A' type) (rpL7/L12); MICLU
G07_orf149	MG093	ribosomal protein L9 (rpL9); BACST
G12_orf161	MG361	ribosomal protein L10 (rpL10); THEMA
G07_orf137	MG081	ribosomal protein L11 (RPL11); THEMA
C12_orf146	MG418	ribosomal protein L13 (rpL13); ECOLI
GT9_orf122	MG161	ribosomal protein L14 (rpL14); BACST
GT9_orf151	MG169	ribosomal protein L15 (rpL15); MYCCA
VXpSPT7_orf139o	MG158	ribosomal protein L16 (rpL16); MYCCA
GT9_orf124a	MG178	ribosomal protein L17 (rpL17); BACSU
GT9_orf116b	MG167	ribosomal protein L18 (rpL18); BACST
K05_orf119	MG444	ribosomal protein L19 (rpL19); BACST
C09_orf127	MG198	ribosomal protein L20 (rpL20); MYCFE
F10_orf100b	MG232	ribosomal protein L21 (rpL21); BACSU
VXpSPT7_orf184	MG156	ribosomal protein L22 (rpL22); HAEIN
VXpSPT7_orf237	MG153	ribosomal protein L23 (rpL23); THEMA
GT9_orf111a	MG162	ribosomal protein L24 (rpL24); BACST
F10_orf104	MG234	ribosomal protein L27 (rpL27); BACSU
C12_orf65	MG426	ribosomal protein L28 (rpL28); BACSU
GT9_orf111b	MG159	ribosomal protein L29 (rpL29); THEMA
H91_orf97	MG257	ribosomal protein L31 (rpL31); ECOLI
G12_orf57	MG363	ribosomal protein L32 (rpL32); HAEIN
P01_orf53	MG325	ribosomal protein L33 (rpL33); BACST
K05_orf48	MG466	ribosomal protein L34 (rpL34); PROMI
C09_orf59	MG197	ribosomal protein L35 (rpL35); BACST
GT9_orf37	MG174	ribosomal protein L36 (rpL36); CHLTR
G07_orf294	MG070	ribosomal protein S2 (rpS2); SPIPL
VXpSPT7_orf273	MG157	ribosomal protein S3 (rpS3); MYCCA
H08_orf205	MG311	ribosomal protein S4 (rpS4); BACSU
GT9_orf219	MG168	ribosomal protein S5 (rpS5); BACSU
G07_orf215	MG090	ribosomal protein S6 (rpS6); ECOLI
G07_orf155	MG088	ribosomal protein S7 (rpS7); BACST
GT9_orf142	MG165	ribosomal protein S8 (rpS8); MYCCA
C12_orf132	MG417	ribosomal protein S9 (rpS9); BACST
VXpSPT7_orf108	MG150	ribosomal protein S10 (rpS10); THEMA
GT9_orf121	MG176	ribosomal protein S11 (rpS11); BACST
G07_orf139	MG087	ribosomal protein S12 (rpS12); BACST
GT9_orf124b	MG175	ribosomal protein S13 (rpS13); BACSU
GT9_orf61	MG164	ribosomal protein S14 (rpS14); MYCCA
C12_orf86	MG424	ribosomal protein S15 (BS18); BACST
K05_orf88	MG446	ribosomal protein S16 (BS17); BACSU
GT9_orf85	MG160	ribosomal protein S17 (rpS17); MYCCA
G07_orf104b	MG092	ribosomal protein S18 (rpS18); ECOLI
VXpSPT7_orf87	MG155	ribosomal protein S19 (rpS19); MYCBO
G12_orf87	MG(M3)	ribosomal protein S20 (rpsT); ECOLI
D12_orf288	MG012	ribosomal protein S6 modification protein (rimK); ECOLI
H91_orf242a	MG252	hypothetical protein YacO (rRNA methylase) homolog; BACSU
VXpSPT7_orf116	MG143	ribosome binding factor A homolog (rbfA); ECOLI
• **Transport and binding proteins - ABC transport [34]**		
A05_orf382	MG303	abc transport ATP-binding protein (artP); ECOLI
D09_orf286a	MG044	spermidine/putrescine transport system permease (potI); ECOLI
D09_orf286b	MG043	spermidine/putrescine transport system permease (potB); HAEIN
D09_orf560L	MG042	spermidine/putrescine transport ATP-binding prot (potA); ECOLI
F10_orf491	MG225	hypothetical protein (gi: 710640) homolog (put. amino acid permease); CLOPE
F10_orf503	MG226	general amino acid permease GAP1 homolog; YEAST
G07_orf376	MG078	oligopeptide transport system permease protein (amiD); STRPN
G07_orf389a	MG077	oligopeptide transport system permease protein (oppB); BACSU
G07_orf423	MG079	oligopeptide transport ATP-binding protein (oppD); BACSU
G07_orf851	MG080	oligopeptide transport ATP-binding protein (oppF); BACSU
GT9_orf303	MG180	histidine transport ATP-binding protein (hisP); ECOLI
R02_orf465	MG065	glutamine transport ATP-binding protein (glnQ); ECOLI
C12_orf225	MG409	phosphate transport system regulatory protein (phoU); ECOLI
C12_orf329	MG410	phosphate transport ATP-binding protein (pstB); ECOLI
C12_orf651V	MG411	phosphate transport system permease protein (pstA); ECOLI
GT9_orf274	MG179	sulfate transport ATP-binding protein (cysA); SYNP
K05_orf284	MG065 (MG467)	sulfate transport ATP-binding protein (cysA); SYNP
A65_orf311	MG121	high affinity ribose transport protein (rbsC); HAEIN
A65_orf572	MG119	hypothetical ABC transporter (yjcW) homolog; ECOLI
E07_orf319	MG189	sn-glycerol-3-phosphate transport system permease protein (ugpE); ECOLI
E07_orf329	MG188	sn-glycerol-3-phosphate transport system permease protein (ugpA); ECOLI
E07_orf586	MG187	sn-glycerol-3-phosphate transport system permease protein (ugpC); ECOLI
A05_orf270L	MG304	abc transport ATP-binding protein (cbiO), SALTY
G07_orf872V	MG071	MG(2+) transport ATPase, P-type 1 (mgtA); ECOLI

Table 1. *Continued*

A05_orf244	MG290	ATP-binding protein P29; MYCHR
A05_orf380V	MG289	high affinity transport system protein P37; MYCHR
A05_orf542	MG291	transport system permease protein P69; MYCHR
D02_orf660	MG390	lactococcin transport ATP-binding protein (lcnDR3); LACLA
D12_orf623	MG014	transport ATP-binding protein (pmd1); SCHPO
D12_orf634	MG015	transport ATP-binding protein (msbA); HAEIN
F10_orf326	MG179	bcrA homolog protein; BACLI
F10_orf750	-	putative ABC transport permease
H08_orf565	MG322	Na(+) translocating ATPase subunit J (ntpJ); ENTHR
K05_orf339	MG467	devA protein homolog; ANASP
• **Transport and binding proteins - PTS transport [7]**		
E09_orf143V	-	PTS system mannitol-specific component IIA (EIIA-MTL)(mtlF); STRMU
E09_orf379	-	PTS system mannitol-specific component IIA (EIIA-MTL)(mtlA); STACA
R02_orf694	MG062	fructose-permease IIBC component (fruA); ECOLI
GT9_orf940o	MG069	PTS system, glucose-specific IIABC component (EIIABC-GLC); BACSU
D09_orf88	MG041	phosphocarrier protein HPr (ptsH); MYCCA
P02_orf159	-	hypothetical phosphotransferase protein YjfU homolog; ECOLI
C12_orf572	MG429	PEP-dependent HPr protein kinase phosphoryltransferase (Enzyme I) (ptsI); STRSL
• **Transport and binding proteins - Other transport systems [3]**		
B01_orf264	MG033	glycerol uptake facilitator (glpF); BACSU
R02_orf564o	MG061	hexosephosphate transport protein (uhpT); SALTY
A05_orf475	MG294	MG294 homolog(put. permease), MYCGE
• **Other categories - Adaptations and atypical conditions [3]**		
K05_orf140	MG454	osmotically inducible protein (osmC); ECOLI
K05_orf270	MG470	soj homolog protein; BACSU
K05_orf263V	MG463	S-adenosylmethionine-6-N',N'-adenosyl (rRNA) dimethytransferase (ksgA); ECOLI
• **Other categories - Other [188]**		
A05_orf102	-	hypothetical 13.2 KD protein homolog (ylxM); BACSU
A05_orf129	MG296	MG296 homolog, MYCGE
A05_orf290	(MG125)	hypothetical protein (YidA) homolog; ECOLI
A05_orf317	MG302	MG302 homolog, MYCGE
A05_orf370	MG295	hypothetical protein (HI0174); HAEIN
A05_orf395	MG306	MG306 homolog, MYCGE
A05_orf982	MG298	P115 protein homolog (SGC3); MYCHR
A19_orf200	MG264	hypothetical protein (HI0890) homolog; HAEIN
A19_orf282	MG265	hypothetical protein (YidA) homolog; ECOLI
A19_orf292	MG263	hypothetical protein (YidA) homolog; ECOLI
A65_orf100	MG134	hypothetical protein YaaK homolog; BACSU
A65_orf117	MG129	MG129 homolog, MYCGE
A65_orf144	MG132	hypothetical protein Hit1 homolog; YEAST
A65_orf145	MG127	hypothetical protein Ygl1 homolog; STRVR
A65_orf166	MG260 (MG185)	MG260 homolog, MYCGE
A65_orf223	MG117	MG117 homolog, MYCGE
A65_orf251b	MG116	MG116 homolog, MYCGE
A65_orf259	MG128	hypothetical protein HI0072 homolog; HAEIN
A65_orf266	MG133	MG133 homolog, MYCGE
A65_orf281	MG125	hypothetical protein (gi: 973220) homolog; ECOLI
A65_orf285	MG135	MG135 homolog, MYCGE
A65_orf377	MG260 (MG185)	MG260 homolog, MYCGE
A65_orf475	MG123	MG123 homolog, MYCGE
A65_orf493	MG130	hypothetical protein Ysr1 homolog; MYCMY
A65_orf517	MG120	MG120 homolog, MYCGE
A65_orf569	MG139	MG139 homolog, MYCGE
B01_orf108	MG029	hypothetical protein (gi: 606093) homolog; ECOLI
B01_orf168	MG027	MG027 homolog, MYCGE
B01_orf186L	MG032	MG032 homolog, MYCGE
B01_orf203	MG028	MG028 homolog, MYCGE
B01_orf338	MG032	MG032 homolog, MYCGE
B01_orf666	MG032	MG032 homolog, MYCGE
B01_orf672	MG032	MG032 homolog, MYCGE
B01_orf673	MG032	MG032 homolog, MYCGE
C09_orf104	MG191	(MG191 homolog, MYCGE)
C09_orf121	MG202	MG202 homolog, MYCGE
C09_orf143b	MG199	MG199 homolog, MYCGE
C09_orf159	MG207	MG207 homolog, MYCGE
C12_orf141	MG427	MG427 homolog, MYCGE
C12_orf172	MG428	MG428 homolog, MYCGE
C12_orf334	MG413 (MG414)	MG413 homolog, MYCGE
C12_orf344	MG415	MG415 homolog, MYCGE
C12_orf385	MG412	MG412 homolog, MYCGE
C12_orf404	MG432	hypothetical protein (yfiB) homolog; SPICI
C12_orf561	MG423	MG423 homolog, MYCGE
C12_orf839	MG422	MG422 homolog, MYCGE
C12_orf997	MG414	MG414 homolog, MYCGE
D02_orf108	MG388	MG388 homolog, MYCGE
D02_orf129	MG389	MG389 homolog, MYCGE
D02_orf135L	MG067 (MG395, MG068)	MG067 homolog, MYCGE
D02_orf140	MG395 (MG068)	MG395 homolog, MYCGE

Table 1. *Continued*

D02_orf150	MG068 (MG395)	MG068 homolog, MYCGE
D02_orf157L	MG395 (MG068)	MG395 homolog, MYCGE
D02_orf225L	MG068 (MG067, MG395)	MG068 homolog, MYCGE
D02_orf265V	MG068 (MG395, MG067)	MG068 homolog, MYCGE
D02_orf346	MG068 (MG395)	MG068 homolog, MYCGE
D02_orf347	MG067 (MG395, MG068)	MG067 homolog, MYCGE
D02_orf353V	MG068 (MG395)	MG068 homolog, MYCGE
D02_orf569	MG397	MG397 homolog, MYCGE
D09_orf125	MG055	MG055 homolog, MYCGE
D09_orf147	MG059	hypothetical protein A43259 homolog; ENTHR
D09_orf178	MG057	hypothetical protein YabF homolog; BACSU
D09_orf276	MG056	hypothetical protein YabC homolog; BACSU
D09_orf451	MG037	pre-B cell enhancing factor homolog (pbeF); HUMAN
D09_orf518	MG096	MG096 homolog, MYCGE
D09_orf632	MG288 (MG096)	MG288 homolog, MYCGE
D12_orf261	MG009	hypothetical protein yabD homolog; BACSU
D12_orf285	MG011	MG011 homolog, MYCGE
E07_orf1113	MG140	MG140 homolog, MYCGE
E07_orf265	MG260 (MG185)	MG260 homolog, MYCGE
E07_orf324	MG190	hypothetical 28K protein (orf4, P1 operon); MYCPN
E07_orf485	MG260 (MG185)	MG260 homolog, MYCGE
E09_orf136	MG441	MG441 homolog, MYCGE
E09_orf204o	-	protein P30, MYCPN
E09_orf287o	MG439	MG439 homolog, MYCGE
E09_orf302	MG440	MG440 homolog, MYCGE
F04_orf154	MG288 (MG096)	MG288 homolog, MYCGE
F04_orf260V	MG288	MG288 homolog, MYCGE
F10_orf100a	MG233	hypothetical protein YsxB homolog; BACSU
F10_orf141b	MG221	hypothetical protein YabB homolog; ECOLI
F10_orf153	MG230	MG230 homolog, MYCGE
F10_orf158	MG236	MG236 homolog, MYCGE
F10_orf291	MG240	MG240 homolog, MYCGE
F10_orf294	MG237	MG237 homolog, MYCGE
F10_orf308	MG222	hypothetical protein YabC homolog; ECOLI
F10_orf419	MG223	MG223 homolog, MYCGE
F10_orf621	MG241	MG241 homolog, MYCGE
F10_orf632o	MG242	MG242 homolog, MYCGE
F10_orf90	MG220	MG220 homolog, MYCGE
F11_orf114	MG267	MG267 homolog, MYCGE
F11_orf122a	MG284	MG284 homolog, MYCGE
F11_orf197	MG286	MG286 homolog, MYCGE
F11_orf218	MG279	MG279 homolog, MYCGE
F11_orf229	MG268	hypothetical protein YaaF homolog; BACSU
F11_orf287	MG280	MG280 homolog, MYCGE
F11_orf346	MG285	MG285 homolog, MYCGE
F11_orf358b	MG269	MG269 homolog, MYCGE
F11_orf582	MG281	MG281 homolog, MYCGE
F11_orf887	MG277	MG277 homolog, MYCGE
G07_orf1030	MG075	protein P100; MYCPN
G07_orf135	MG074	MG074 homolog, MYCGE
G07_orf138	MG076	MG076 homolog, MYCGE
G07_orf289	MG084	hypothetical protein (yacA) homolog; BACSU
G07_orf312	MG085	MG085 homolog, MYCGE
G07_orf417	MG288 (MG096)	MG288 homolog, MYCGE
G07_orf478o	MG100	PET112 protein homolog; YEAST
G07_orf478V	MG099	amidase homolog (S47454); YEAST
G07_orf479	MG098	MG098 homolog, MYCGE
G12_orf104	MG376	MG376 homolog, MYCGE
G12_orf109	MG353	MG353 homolog, MYCGE
G12_orf136	MG354	MG354 homolog, MYCGE
G12_orf166a	MG342	MG342 homolog, MYCGE
G12_orf166b	MG346	hypothetical protein Ygl3 homolog; BACST
G12_orf210V	MG347	hypothetical protein HI0340 homolog; HAEIN
G12_orf218	MG364	MG364 homolog, MYCGE
G12_orf269	MG374	MG374 homolog, MYCGE
G12_orf281	MG373	MG373 homolog, MYCGE
G12_orf325	MG371	hypothetical 28K protein (P1 operon) homolog; MYCPN
G12_orf326	MG370	hypothetical protein (HI0176) homolog; HAEIN
G12_orf328b	MG350	MG350 homolog, MYCGE
G12_orf348	MG343	MG343 homolog, MYCGE
G12_orf387	MG372	MG372 homolog, MYCGE
G12_orf413	MG349	MG349 homolog, MYCGE
G12_orf558	MG369	MG369 homolog, MYCGE
G12_orf664	MG366	MG366 homolog, MYCGE
GT9_orf148	MG260	MG260 homolog, MYCGE
GT9_orf434	MG181	MG181 homolog, MYCGE
H03_orf235	MG381	MG381 homolog, MYCGE
H08_orf157b	MG321	MG321 homolog, MYCGE
H08_orf193	MG319	MG319 homolog, MYCGE
H08_orf231	MG323	hypothetical protein YZAC homolog; BACSU
H08_orf263	MG313	MG313 homolog, MYCGE
H08_orf287	MG320	(cytochrome C oxidase polypeptide I (CtaD); BACSU)
H08_orf314	MG315	MG315 homolog, MYCGE
H08_orf345	MG307	MG307 homolog, MYCGE

Table 1. *Continued*

H08_orf369	MG316	(competence locus E (comE3); BACSU)
H08_orf448	MG314	MG314 homolog, MYCGE
H08_orf572o	MG307	MG307 homolog, MYCGE
H08_orf591	MG321	MG321 homolog, MYCGE
H08_orf726	MG307	MG307 homolog, MYCGE
H10_orf149	MG211	MG211 homolog, MYCGE
H10_orf196	MG208	MG208 homolog, MYCGE
H10_orf208	MG214	hypothetical protein P35155 homolog; BACSU
H10_orf309	MG209	hypothetical protein YceC homolog; ECOLI
H91_orf213	MG248	MG248 homolog; MYCGE
H91_orf224	MG243	MG243 homolog, MYCGE
H91_orf239	MG247	hypothetical protein YgiH homolog; ECOLI
H91_orf258	MG256	MG256 homolog, MYCGE
H91_orf281	MG246	MG246 homolog, MYCGE
H91_orf534	MG255	MG255 homolog, MYCGE
H91_orf677	MG260	MG260 homolog, MYCGE
K04_orf202	MG105	MG105 homolog, MYCGE
K04_orf222	MG101	MG101 homolog, MYCGE
K04_orf278L	MG110	hypothetical protein YjeQ homolog; ECOLI
K04_orf280	MG103	MG103 homolog, MYCGE
K05_orf169	MG459	hypothetical protein HI0671 homolog; HAEIN
K05_orf234	MG449	MG449 homolog, MYCGE
K05_orf237	MG450	degV homolog protein; BACSU
K05_orf251	MG452	MG452 homolog, MYCGE
K05_orf271	MG442	MG442 homolog, MYCGE
K05_orf345	MG456	MG456 homolog, MYCGE
K05_orf385	MG464	hypothetical protein 1 (S42122); MYCCA
K05_orf401	MG443	hypothetical protein (P27712); SPICI
K05_orf425	MG461	MG461 homolog, MYCGE
K05_orf499	MG447	MG447 homolog, MYCGE
P01_orf1033	MG328	MG328 homolog, MYCGE
P01_orf197	MG333	hypothetical protein HI1366 homolog; HAEIN
P01_orf209	MG331	MG331 homolog, MYCGE
P01_orf235	MG332	hypothetical protein HI0315 homolog; HAEIN
P01_orf293	MG326	degV homolog protein; BACSU
P01_orf341	marginal MG025	hypothetical protein YibD homolog; ECOLI
P02_orf140	MG337	MG337 homolog, MYCGE
P02_orf218	-	hypothetical protein YjfV homolog; ECOLI
P02_orf305	-	hypothetical protein YjfW homolog; ECOLI
P02_orf316	MG338	MG338 homolog, MYCGE
P02_orf408	MG336	nitrogen fixation protein (nifS); HAEIN
P02_orf427	MG288 (MG096)	MG288 homolog, MYCGE
P02_orf458	MG096 (MG288)	MG096 homolog, MYCGE
P02_orf509	MG288 (MG096)	MG288 homolog, MYCGE
P02_orf660	-	hypothetical protein YjfS homolog; ECOLI
R02_orf1386V	MG064	MG064 homolog, MYCGE
R02_orf147	MG260	MG260 homolog, MYCGE
R02_orf469	MG061	MG061 homolog, MYCGE
R02_orf524	MG068 (MG067)	MG068 homolog, MYCGE
VXpSPT7_orf269	MG145	hypothetical protein (YaaC) homolog; PSEFL
VXpSPT7_orf377	MG147	MG147 homolog, MYCGE
VXpSPT7_orf402	MG144	MG144 homolog, MYCGE
VXpSPT7_orf445	MG148	MG148 homolog, MYCGE
• no classification so far [86]		
A19_orf1140	-	-
A19_orf129	-	-
A19_orf204	-	-
A19_orf229V	-	-
A19_orf591	-	-
A65_orf115	-	-
A65_orf118	-	-
B01_orf103b	-	-
B01_orf116L	-	-
B01_orf147	-	-
b01_orf182l	-	-
B01_orf274	-	-
C09_orf130b	-	-
C09_orf140o	-	-
C09_orf165	-	-
C09_orf172	-	-
C09_orf223	-	-
C09_orf251	-	-
C09_orf404	-	-
C09_orf422	-	-
C09_orf718	-	-
C12_orf181o	-	-
C12_orf247	-	-
D02_orf100	-	-
D02_orf109	-	-
D02_orf122a	-	-
D02_orf122b	-	-
D02_orf128	-	-
D09_orf127a	-	-

Table 1. *Continued*

D12_orf131	-	-
D12_orf235	-	-
D12_orf257	-	-
E07_orf133	-	-
E07_orf140	-	-
E07_orf163	-	-
E07_orf166	-	-
E07_orf175	-	-
E07_orf179	-	-
E07_orf228	-	-
E09_orf136L	marginal MG440	-
E30_orf352	-	-
F04_orf120	-	-
F04_orf150	-	-
F10_orf218	-	-
F10_orf357	marginal MG011	-
F10_orf565	-	-
F10_orf741	-	-
F11_orf148o	-	-
F11_orf879	-	-
G12_orf140b	-	-
G12_orf168	-	-
G12_orf225	-	-
GT9_orf113	-	-
H03_orf152	-	-
H08_orf102	-	-
H10_orf119	-	-
H10_orf206	-	-
H10_orf220L	-	-
H91_orf115	-	-
H91_orf180	-	-
H91_orf216	-	-
K05_orf101a	-	-
K05_orf106	-	-
K05_orf1882	marginal MG064	-
K05_orf250	-	-
P01_orf140	-	-
P01_orf199	-	-
P01_orf243	-	-
P02_orf103b	-	-
P02_orf126	-	-
P02_orf143	-	-
P02_orf147	-	-
P02_orf163	-	-
P02_orf196	-	-
P02_orf253	-	-
P02_orf474	-	-
R02_orf101	-	-
R02_orf105	-	-
R02_orf140	-	-
R02_orf150	-	-
R02_orf183o	-	-
R02_orf254	-	-
R02_orf264	-	-
R02_orf329	-	-
R02_orf440	-	-
VXpSPT7_orf112	-	-
• hypothetical ORFs derived from repetitve DNA elements [46]		
A05_orf139	-	-
A19_orf211	-	-
A65_orf115	-	-
B01_orf147	-	-
C09_orf140o	-	-
C09_orf149a	-	-
E07_orf163	-	-
F11_orf148o	-	-
G12_orf168	-	-
H08_orf157a	marginal MG321	-
H91_orf180	-	-
P01_orf199	-	-
P02_orf103b	-	-
P02_orf196	-	-
R02_orf138	-	-
R02_orf140	-	-
R02_orf183o	-	-
C09_orf149b	-	adhesin P1 (group 2) homolog; MYCPN
H08_orf329V	MG321	adhesin P1 (group 2) homolog; MYCPN
A65_orf465V	MG191	adhesin P1 (group 2) homolog; MYCPN
E07_orf413	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
E07_orf256L	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
A05_orf278	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
H08_orf270	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
P02_orf422V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN

Table 1. *Continued*

P02_orf527V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
F11_orf533L	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
P01_orf208V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
GT9_orf438V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
GT9_orf127	-	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
GT9_orf313	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
C09_orf428V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
A19_orf737V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
E07_orf221V	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
R02_orf347L	MG191	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
G12_orf325	MG371	hypothetical 28K protein (P1 operon) homolog; MYCPN
E07_orf224	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
E07_orf434	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
C09_orf272	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
A05_orf493	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
R02_orf301	-	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
R02_orf173	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
H08_orf445	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
P02_orf381	(MG192)	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
H91_orf322	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
H91_orf272	MG192	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN

- **RNA - rRNA [3]**

5S rRNA
16S rRNA
23S rRNA

- **RNA - tRNA [33 tRNAs in 14 genes/operons]**

Arg-tRNA gene (CGA); MYCPN
Arg-tRNA gene (CGC); MYCPN
Arg-tRNAgene (AGA); MYCPN
Asn-tRNA(AAC), Glu-tRNA(GAA), Thr-tRNA(ACG), Val-tRNA(GTA), Thr-tRNA(ACA), Lys-tRNA(AAG), Leu-tRNA(CTA) genes; MYCPN
Cys-tRNA(TGC), Pro-tRNA(CCA), Met-tRNA(ATG), Ile-tRNA(ATG), Ser-tRNA(TCA), fMet-tRNA(ATG), Asp-tRNA(GAC) and Phe-tRNA(TTC) genes; MYCPN
Gly-tRNA(GGC) gene; MYCPN
His-tRNA(CAC) gene; MYCPN
Ile-tRNA(ATC), Ala-tRNA(GCA) genes; MYCPN
Thr-tRNA(GGU) gene; MYCPN
Ser-tRNA (AGC) gene; MYCPN
Ser-tRNA(TCC), Ser-tRNA(TCG) genes; MYCPN
Trp-tRNA (TGA)gene; MYCPN
Trp-tRNA(TGG) gene; MYCPN
Tyr-tRNA (TAC), Glu-tRNA (CAA), Lys-tRNA (AAA), Leu-tRNA (TTA), Gly-tRNA (GGA) genes; MYCPN

- **RNA - other [3]**

4.5S RNA; MYCPN
10sa RNA; MYCGE
RNaseP RNA; MYCGE

MG is the name of the corresponding ORF in *M.genitalium* (9).

coding densities have been also estimated for the smaller *M.genitalium* genome (9) and for the genome of *Haemophilus influenzae* which is more than twice as large (30). The length of the proposed proteins in *M.pneumoniae* ranges from 37 (4.3 kDa) to 1882 (209.4 kDa) amino acids (Fig. 3). One of the largest proteins is the cytadherence accessory protein HMW2 (F10_orf1818) and the smallest identified protein is the 37 amino acid ribosomal protein L36 (GT9_orf37). For practical reasons we introduced at the beginning of the sequence analysis a cut-off point of 100 amino acids for proposed proteins unless we found smaller proteins such as some of the ribosomal proteins during the initial BLASTX homology search. All intergenic or non coding regions were reanalyzed with a cut-off point of 50 amino acids and searches were done for specific small proteins. However, we cannot exclude the possibility that some of the smaller proteins, not showing similarities to known proteins from other organisms, have been missed in our analysis.

The codon usage of *M.pneumoniae* is summarized in Table 3. We compared it for all proposed genes, for the subsets of genes with a low G+C (content below 35 mol%) and high G+C content (between 50 and 56 mol%) and for all 50 ribosomal protein genes (42.8 mol%) as an example for frequently translated genes. Codon usage of the low and high G+C content subfractions is clearly influenced by the DNA composition, favouring either codons with G/C or A/T at the third position. The codon usage pattern differs also for the complete genome and for genes which are frequently expressed like the ones coding for ribosomal proteins.

The most frequently used codons are AUU (Ile, 4.6%); AAA (Lys, 4.6%); UUU (Phe, 4.3%); GAA (Glu, 4.2%) and UUA (Leu, 3.9%) and the most common amino acids are Leu (10.3%), Lys (8.5%), Ile (6.6%), Ala (6.6%) and Val (6.5%). The high value for Lys is in agreement with the relative high percentage of proposed proteins with calculated isoelectric points between pH 9 and 12 (Fig. 4). The least frequently used codons are UGC (Cys, 0.2%); CGA (Arg, 0.25%); AGG (Arg, 0.29%); AGA (Arg, 0.4%) and UGU (Cys, 0.55%).

All *M.pneumoniae* gene products were classified (Table 1 and 2), with some minor modifications, in accordance with criteria introduced for *Escherichia coli* (31) and adapted for the classification of putative genes from *H.influenzae*. We added

Table 2. Summary of the functional classification of the ORFs

• Biosynthesis of cofactors, prosthetic groups and carrier	8
Folic acid	5
Heme and porphyrin	1
Thioredoxin	2
• Cell envelope	54
Membranes, lipoproteins and porines	42
Surface structures and cytadherence	8
Surfaces polysaccharides, lipopolysaccharides and antigens	4
• Cellular processes	20
Cell division	2
Cell killing	1
Chaperones	7
Detoxification	1
Protein and peptide secretion	9
• Central intermediary metabolism	6
Other	5
Phosphorous compounds	1
• Energy metabolism	39
Aerobic	3
Amino acids and amines	5
Anaerobic	1
ATP-proton motive force interconversion	9
Glycolysis	10
Pentose Phosphate pathway	2
Pyruvate DHase	4
Sugars	5
• Fatty acid and phospholipid metabolism	9
• Purines, pyrimidines, nucleosides and nucleotides	18
2'-Deoyribonucleotide metabolism	3
Nucleotide and nucleoside interconversions	2
Purine ribonucleotide biosynthesis	3
Salvage of nucleosides and nucleotides	8
Sugar-nucleotide biosynthesis and conversions	2
• Pyridine nucleotide metabolism	1
• Regulatory function	8
• Replication	46
DNA replication, restriction, modification, recombination and repair	46
• Transcription	13
Degradation of RNA	2
RNA synthesis, modification and DNA transcription	11
• Translation	99
Amino acyl tRNA synthetases and tRNA modification	24
Degradation of proteins, peptides and glycopeptides	8
Protein modification and translation factors	15
Ribosomal proteins: synthesis and modification	52
• Transport and binding proteins	44
ABC transport	34
PTS transport	7
Other transport systems	3
• Other categories	191
Adaptations and atypical conditions	3
Other	188
• hypothetical ORFs derived from repetitive DNA elements	46
• no classification so far	86
• RNA	39
rRNA	3
tRNA	33
other	3

'cytadherence associated proteins' to the category of cell envelope–surface structures, since evidence is mounting, that *M.pneumoniae* possesses a cytoskeleton-like organization which stabilizes the bacterium and protects it against osmotic lysis (2). The category of transport and binding proteins was altered by subdivision into three groups namely, into PTS-, ABC- and other transport systems. To facilitate the orientation on the gene map we added a list which contains all proposed ORFs and RNAs in nummerical order (Table 4).

More details on this very general analysis will be made public on the www (http://www.zmbh.uni-heidelberg.de/M_pneumoniae).

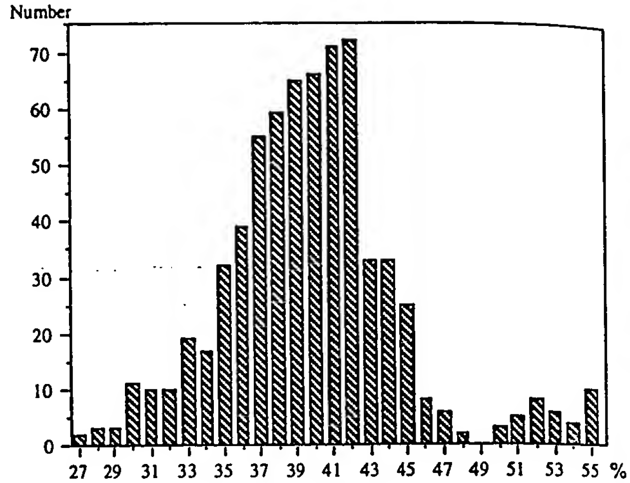


Figure 2. Distribution of the G+C content of the coding sequences of all *M.pneumoniae* ORFs.

DNA replication and repair

The central enzyme for DNA replication in bacteria is the DNA polymerase III holoenzyme (32), which consists of 10 subunits in *E.coli*, a DNA polymerase subunit α and nine accessory proteins (ε, υ, τ, γ, δ, δ′, χ, ψ and β). *Mycoplasma pneumoniae* codes for two potential α subunits (the gene name in the literature is either dnaE or polC). Both proposed α subunits, A19_orf872 and B01_orf1443, differ in length and also in their degree of similarity to the α subunits from *E.coli* and *Bacillus subtilis*. The protein from B01_orf1443 shares the highest similarity with the α subunit from Gram-positive bacteria including the motif for a 3′–5′ exonuclease activity which is typical for these bacteria. In contrast, the orf A19_orf872 is most similar to the α subunit from *E.coli* and does not contain a 3′-5′ exonuclease domain. The 3′–5′ exonuclease activity in *E.coli* is encoded by a separate gene (dnaQ), which has not been found in *M.pneumoniae*. Of the other subunits which build the DNA polymerase III holoenzyme in *E.coli* (32) only the subunits β (dnaN), δ′(holB), γ and τ (dnaX) are present in *M.pneumoniae*, indicating a simplified replication complex compared with the Gram-negative bacteria *E.coli* and *H.influenzae*. Presently, it cannot be excluded that other proteins replace these subunits in *M.pneumoniae*. A true comparison with a phylogenetically closer related Gram-positive bacterium like *B.subtilis* is not possible since the *Bacillus* DNA polymerase III holoenzyme complex has not been defined as yet and the nucleotide sequence of the entire *B.subtilis* genome has not been completed.

Mycoplasma pneumoniae does not code for a DNA polymerase I (polA)-like DNA repair enzyme. Instead, we find a truncated polA gene (A19_orf291) comprising only the 5′–3′ exonuclease domain, whereas in *E.coli* and *B.subtilis* the polA gene is much larger and codes for the 5′–3′ exonuclease and a 5′–3′ polymerase-specific domain.

Experimental results on DNA polymerase enzymatic activities in mycoplasmas are confusing. It was claimed that the DNA polymerase III of *Mollicutes* lacks the 3′–5′ exonuclease proofreading activity in general (33) and this was taken as an explanation for the observed genetic instability of many *Mollicutes* species (4). Recently, the nucleotide sequence of the polC gene of

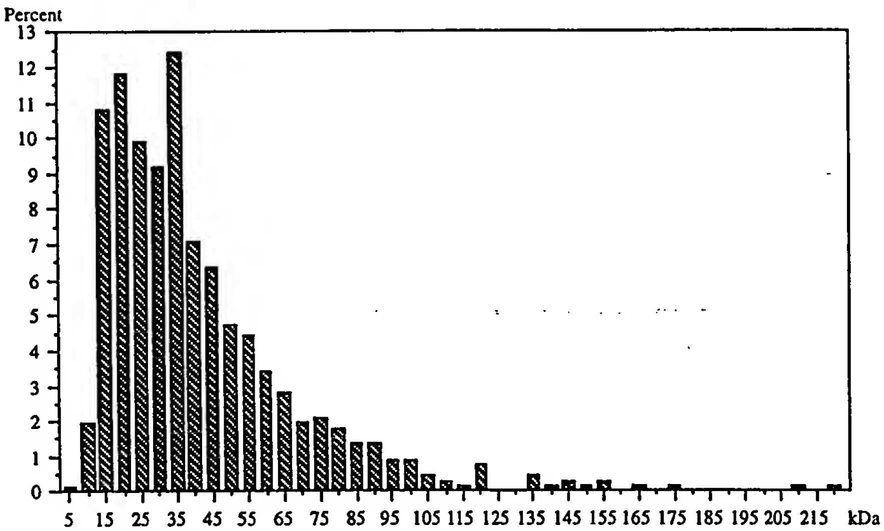


Figure 3. Distribution of all *M.pneumoniae* proteins according to their molecular weight.

Mycoplasma pulmonis and experimental results on enzyme purification and characterization of enzyme activities were published (34). The results indicated that the polC gene from *M.pulmonis* also codes for a 3′–5′ exonuclease, and that the size of the predicted PolC protein, 1435 amino acids, is very similar to the PolC homolog B01_orf1443 in *M.pneumoniae* and that the polymerase could be inhibited by compounds specific for PolC proteins of Gram-positive bacteria. Furthermore, the authors provided some experimental evidence for a second, smaller enzyme with DNA polymerase activity. Considering the characterization data of DNA polymerase activities in *M.pulmonis* and the nucleotide sequence data on DNA polymerase genes of *M.pneumoniae* and *M.genitalium* (9,35), one can conclude that at least these three *Mycoplasma* species have two DNA polymerase (polC) genes coding for a larger protein (≈1400 amino acids) with a 3′–5′ exonuclease activity and with the highest sequence similarities to the Gram-positive *B.subtilis* polymerase III. Therefore it is unlikely that an increased mutation frequency is caused by the DNA replication process. The nucleotide sequence of the smaller Pol III homolog (≈100 kDa) of *M.pneumoniae* and *M.genitalium* (9,35) resembles more the polC gene from the Gram-negative *E.coli*. This is also emphasized by the absence of the 3′–5′ exonuclease domain in the proposed genes. The gene for the smaller, Gram-negative typical PolC has not yet been found in *M.pulmonis*, but during the purification of the larger PolC, a second polymerase activity lacking exonuclease activity has been identified. The function of the exonuclease negative DNA polymerase can only be elucidated experimentally and it remains to be seen if it can substitute for the function of the polymerase I (PolA) in combination with the proposed 5′–3′ exonuclease of the truncated polA gene (A19_orf291). This topic has been also discussed for *M.genitalium* (35).

In addition to the DNA polymerase many more gene products are necessary for DNA replication, e.g. initiation, elongation and termination (32). The most obvious functions missing in *M.pneumoniae* according to the sequence analysis are an RNaseH for primer removal and a protein for the termination of replication.

The number of genes involved in DNA repair is considerably smaller in *M.pneumoniae* than in the 'standard' eubacteria *E.coli* and *B.subtilis* or even *H.influenzae* with the smaller genome.

Mycoplasma pneumoniae codes only for 13 of the genes known to be involved in excision repair of DNA, recombination and SOS repair. Thus the genes recB, recC, recD, recG and ruvC involved in recombination are missing as well as the genes recN, recO, recQ and recR involved in SOS repair in *E.coli*. Nevertheless, a rudimentary stock of enzymes has been conserved in *M.pneumoniae* to permit homologous recombination [RecA, Ssb, PolA (see above), GyrA, GyrB, RuvA and RuvB] (36), excision repair (37) and a kind of truncated SOS repair (38). In particular missing is the lexA gene which plays a central role in regulating the SOS response including the expression of the recA gene in other bacteria.

We were also unable to find components of the so called mismatch-repair system encoded by the mutS, mutL and mutH genes. Since bacteria which normally carry the mut genes show a reduced genetic stability, if these genes are mutated, it seems likely that the absence of these genes in mycoplasmas causes an increased mutation rate (65).

Transcription

The DNA dependent RNA polymerase of *M.pneumoniae* is coded by the conserved genes rpoA (α subunit), rpoB (β subunit), rpoC (β' subunit) and rpoE (δ' subunit). The only sigma factor found (H91_orf499) shares the highest similarity with the sigma factor SigA from *B.subtilis* (39). Presently, not enough experimental data are available for defining promoter sequences in *M.pneumoniae*. The promoter of only three genes/operons have been determined experimentally by primer extension. These genes are the P1 operon (14), the ribosomal RNA operon (40) and F10_orf405 (27). The –10 region and to a lesser extent the –35 region of these three examples are comparable with consensus promoters sequences in *B.subtilis* (41). Termination of transcription seems to be independent of the termination factor Rho, since the corresponding gene could not be found. Transcription stops on typical terminator sequences which are short interrupted palin-

Table 3. Codon usage of different sets of *M.pneumoniae* ORFs: all 677 ORFs; ORFs with a G+C content <35 mol%; codon usage of the adhesin P1 and ORF6 (high G+C content); ribosomal ORFs as examples for frequently expressed proteins

AmAcid	Codon	all MP ORFs (677) /1000	GC<35% /1000	high GC (P1+orf6) /1000	ribosomal ORFs /1000
Ala	GCA	13.76	14.92	8.43	14.90
Ala	GCC	16.50	8.09	27.75	16.95
Ala	GCG	11.05	4.43	22.48	13.12
Ala	GCT	25.20	22.80	25.64	30.62
Arg	AGA	4.02	11.22	2.46	5.19
Arg	AGG	2.84	3.70	4.21	1.37
Arg	CGA	2.48	3.55	2.81	3.42
Arg	CGC	10.72	4.59	14.75	22.83
Arg	CGG	5.00	0.94	5.27	8.20
Arg	CGT	9.68	5.63	6.32	21.46
Asn	AAC	37.01	27.91	41.80	41.69
Asn	AAT	25.09	45.50	24.24	15.72
Asp	GAC	19.16	13.88	25.99	14.63
Asp	GAT	30.40	39.18	32.31	19.68
Cys	TGC	2.09	2.82	0.00	2.32
Cys	TGT	5.39	5.48	0.00	3.96
Gln	CAA	37.90	39.55	31.96	35.95
Gln	CAG	15.65	7.46	21.07	8.34
Glu	GAA	42.01	53.22	20.02	39.64
Glu	GAG	14.71	12.47	12.29	11.34
Gly	GGA	6.38	9.29	8.43	7.52
Gly	GGC	11.81	9.34	22.13	12.17
Gly	GGG	8.95	2.30	25.99	8.61
Gly	GGT	27.90	22.33	27.75	34.86
His	CAC	11.86	6.16	8.08	16.54
His	CAT	6.17	6.16	2.81	4.24
Ile	ATA	5.46	12.84	1.40	1.78
Ile	ATC	14.39	13.10	11.59	13.94
Ile	ATT	45.99	48.21	16.16	47.57
Leu	CTA	10.62	10.64	3.86	8.88
Leu	CTC	12.23	6.47	26.69	13.81
Leu	CTG	9.54	5.17	10.89	6.01
Leu	CTT	10.06	18.10	8.78	7.38
Leu	TTA	39.24	46.54	19.32	34.03
Leu	TTG	21.48	17.48	22.48	16.54
Lys	AAA	46.27	73.20	24.24	61.92
Lys	AAG	39.08	29.84	33.02	63.01
Met	ATG	15.60	13.98	7.38	21.32
Phe	TTC	12.75	16.23	10.89	7.52
Phe	TTT	43.03	53.17	25.64	24.06
Pro	CCA	10.86	9.76	16.51	12.03
Pro	CCC	9.05	3.13	23.18	7.11
Pro	CCG	6.65	2.40	14.05	7.52
Pro	CCT	8.30	9.86	9.13	9.16
Ser	AGC	10.62	10.49	11.94	8.20
Ser	AGT	21.04	21.76	28.10	12.85
Ser	TCA	8.74	13.20	8.43	8.61
Ser	TCC	9.59	6.73	22.48	9.84
Ser	TCG	6.43	3.18	15.10	5.06
Ser	TCT	8.16	15.03	5.97	6.15
Thr	ACA	10.38	15.18	8.43	8.47
Thr	ACC	21.92	11.74	45.66	27.88
Thr	ACG	7.90	3.60	18.97	6.56
Thr	ACT	19.32	24.16	10.89	17.22
Trp	TGA	6.06	8.77	9.83	2.32
Trp	TGG	5.82	3.60	9.13	4.10
Tyr	TAC	17.94	15.34	16.51	13.67
Tyr	TAT	14.26	20.04	10.89	9.16
Val	GTA	13.73	11.64	7.73	21.05
Val	GTC	11.03	4.85	15.45	8.47
Val	GTG	18.73	6.37	29.50	21.46
Val	GTT	21.17	27.50	14.05	23.10
xxx	TAA	2.05	2.97	0.35	1.91
xxx	TAG	0.78	0.83	0.35	5.06

dromic regions followed by a run of U residues. The Nus transcription termination factors, of which NusA (E07_orf540) and NusG (D09_orf320) are present, may play a role in the termination of transcription. NusB and NusC are absent. NusA is involved in termination and NusG in antitermination in other bacteria. Finally, GreA promotes elongation by the RNA polymerase by utilizing a novel transcript-cleavage reaction (42).

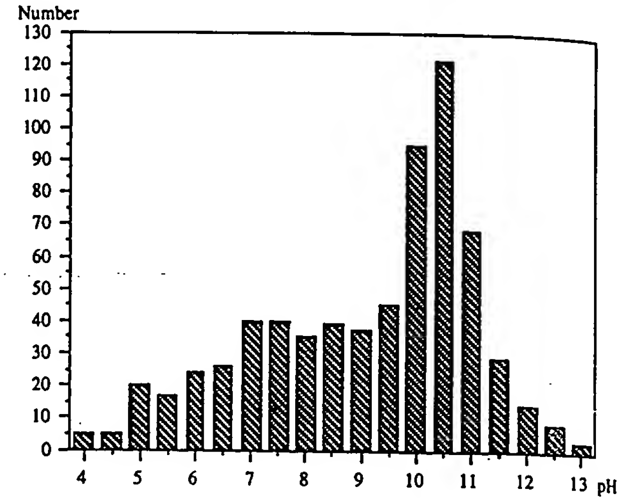


Figure 4. Distribution of all *M.pneumoniae* proteins according to their predicted isoelectric point (IP).

Gene expression and regulation

Regulation of gene expression in *M.pneumoniae* has not been studied so far. Therefore we do not know how this bacterium coordinates the synthesis of those gene products which are essential for reproduction. Also, *M.pneumoniae* has to sense and respond to environmental changes. This requires a signal transduction system. The presence of only one sigma factor (sigA, H91_orf499) which is also the only one of all proposed proteins showing the characteristic helix–turn–helix (HTH) motif, suggests that the response to external stimuli is not controlled by the level of expression of alternative sigma factors.

The presence of a *cis*-acting conserved palindromic repeated sequence in front of four heat shock genes, similar to the 'CIRCE' element first identified in *B.subtilis* (43) and the identification of the proposed repressor (C09_orf351, hrcA), indicates that the heat shock response in *M.pneumoniae* is regulated by the interaction of this repressor with the CIRCE element, and provides an example for a negative regulation of gene expression in *M.pneumoniae*.

The two-component signal transduction system (44), consisting of a sensor and a response regulator, which has been found in many prokaryotic and eukaryotic organisms is believed to be essential for all cells. Nevertheless, based on sequence similarity we were unable to detect any such system in *M.pneumoniae*.

Concerning other proteins with regulatory functions we identified several GTP-binding proteins and other proteins like the virulence associated protein vacB (K04-orf726). These regulatory proteins act by unknown mechanisms.

Translation

The translation machinery of *M.pneumoniae* is rather extensive. About 15% of all proposed ORFs, are involved in translation including 19 tRNA synthetases, 50 ribosomal proteins, various factors and enzymes, 33 tRNAs, one ribosomal RNA operon with one copy of each 5S, 16S and 23S rRNA (45), and a gene coding for the 10Sa RNA. The conservation of the 10Sa RNA which functions as tRNA and mRNA and is implicated in *trans*-translation (66), is interesting in evolutionary terms. Three exceptions are

Table 4. List of the proposed ORFs, RNAs and REPs in numerical order starting with E07_orf540o on the gene map (Fig. 1)

Number	Genome Position	Name	Annotation
001	663**815435 (rc!)	E07_orf540o	N-utilization substance protein A homolog (nusA); BACSU
002	4081..740	E07_orf1113	MG140 homolog, MYCGE
003	6641..4257	E07_orf794	putative lipoprotein, MG260 homolog, MYCGE
004	7325..6924	E07_orf133	-
005	8482..7808	E07_orf224	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
	8620..7896	REPMP5	repetitive DNA sequence REPMP5
006	9614..8310	E07_orf434	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
007	10589..10167	E07_orf140	-
008	12589..11132	E07_orf485	MG260 homolog, MYCGE
009	13393..12596	E07_orf265	MG260 homolog, MYCGE
010	14250..13711	E07_orf179	-
011	15843..14602	E07_orf413	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	16274..14754	REPMP2/3	repetitive DNA sequence REPMP2/3
012	16944..16417	E07_orf175	-
013	20717..17061	E07_orf1218	hypothetical 130K protein (orf6; P1 operon); MYCPN -
	20717..18017	REPMP5	repetitive DNA sequence REPMP5
	23560..21760	REPMP2/3	repetitive DNA sequence REPMP2/3
014	25606..20723	E07_orf1627	ADP1_MYCPN adhesin P1 (orf5, P1 operon); MYCPN
	25606..24060	REPMP4	repetitive DNA sequence REPMP4
015	26593..25619	E07_orf324	hypothetical 28K protein (orf4, P1 operon); MYCPN
	26823..27091	REPMP1	repetitive DNA sequence REPMP1
016	26844..27335	E07_orf163	-
017	27572..28072	E07_orf166	-
018	28321..29007	E07_orf228	-
019	30544..29585	E07_orf319	sn-glycerol-3-phosphate transport system permease protein (ugpE); ECOLI
020	31505..30516	E07_orf329	sn-glycerol-3-phosphate transport system permease protein (ugpA); ECOLI
021	33258..31498	E07_orf586	sn-glycerol-3-phosphate transport system permease protein (ugpC); ECOLI
022	34187..33282	E07_orf301	putative lipoprotein, MG186 homolog, MYCGE
	35192..36457	REPMP2/3	repetitive DNA sequence REPMP2/3
023	35415..34645	E07_orf256L	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
024	36396..35731	E07_orf221V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	37389..37148	REPMP1	repetitive DNA sequence REPMP1
025	37422..37000	C09_orf140o	-
	38383..37821	REPMP2/3	repetitive DNA sequence REPMP2/3
026	38832..38383	C09_orf149b	adhesin P1 (group 2) homolog; MYCPN
027	39981..39532	C09_orf149a	-
	40650..39538	REPMP4	repetitive DNA sequence REPMP4
028	41980..41438	C09_orf180	-
029	42851..42372	C09_orf159	MG207 homolog, MYCGE
030	44647..42887	C09_orf586L	excinuclease ABC subunit C (uvrC), BACSU
031	44679..45734	C09_orf351	protein (hrcA) homolog, BACSU
032	48090..45721	C09_orf789	topoisomerase IV subunit A (parC), BACSU
033	49997..48090	C09_orf635	topoisomerase IV subunit B (parE), BACSU
	50032..50105	mptgt	Thr-tRNA(GGU) gene; MYCPN
034	50488..50123	C09_orf121	MG202 homolog, MYCGE
035	51141..50488	C09_orf217	heat shock protein GrpE, HAEIN
036	53896..51164	C09_orf910	DnaJ homolog protein, MYCCA
037	54231..54662	C09_orf143b	MG199 homolog, MYCGE
038	55020.. 54637	C09_orf127	ribosomal protein L20 (rpl20); MYCFE
039	55210..55031	C09_orf59	ribosomal protein L35 (rpL35); BACST
040	55821..55216	C09_orf201	translation initiation factor IF3 (infC); MYCFE
041	57713..55911	C09_orf600	carnitine palmitoyltransferase II precursor(cpt2); HUMAN
042	58374..57703	C09_orf223	-
043	59315..58923	C09_orf130b	-
044	61443..60175	C09_orf422	-
045	64103..61947	C09_orf718	-
046	64524..64027	C09_orf165	-
047	66418..65204	C09_orf404	-
048	67175..66420	C09_orf251	-
049	69705..67288	C09_orf805	phenylalanyl-tRNA synthetase beta chain (pheT); BACSU
050	70733..69708	C09_orf341	phenylalanyl-tRNA synthetase alpha-subunit (pheS); BACSU
051	71881..71567	C09_orf104	(MG191 homolog, MYCGE)
052	71891..72409	C09_orf172	-
053	73896..73078	C09_orf272	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
	74668..72883	REPMP5	repetitive DNA sequence REPMP5
054	75998..74712	C09_orf428V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	76039..74736	REPMP4	repetitive DNA sequence REPMP4
	76973..76691	REPMP1	repetitive DNA sequence REPMP1
055	77006..76455	R02_orf183o	-
056	78388..77345	R02_orf347L	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	79072..77697	REPMP2/3	repetitive DNA sequence REPMP2/3
057	79517..79074	R02_orf147	MG260 homolog, MYCGE
058	81440..79815	R02_orf541	putative lipoprotein, MG260 homolog, MYCGE
059	82410..81616	R02_orf264	-
060	83174..82410	R02_orf254	-
	83460..83358	5s rRNA	5S rRNA
	86408..83682	23s rRNA	23S rRNA
	88155..86632	16s rRNA	16S rRNA
061	90177..89755	R02_orf140	-
	90202..89903	REPMP1	repetitive DNA sequence REPMP1
062	91516..90611	R02_orf301	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
063	91892..91371	R02_orf173	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
064	92626..92210	R02_orf138	-
	92692..90643	REPMP5	repetitive DNA sequence REPMP5
065	93692..92703	R02_orf329	-
066	94854..93847	R02_orf335	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
067	95651..95346	R02_orf101	-
068	97118..96666	R02_orf150	-
069	97607..97290	R02_orf105	-
070	99191..97869	R02_orf440	-
071	100872..99298	R02_orf524	MG068 homolog, MYCGE
072	102523..100922	R02_orf533	putative lipoprotein, MG067 homolog, MYCGE
073	104479..102533	R02_orf648	transketolase 1 (TK 1; tktB); RHOSH
074	105897..104500	R02_orf465	glutamine transport ATP-binding protein (glnQ); ECOLI
075	110057..105897	R02_orf1386V	MG064 homolog, MYCGE
076	111196..110294	R02_orf300	1-phosphofructokinase (fruK); HAEIN
077	113273..111189	R02_orf694	fructose-permease IIBC component (fruA); ECOLI
	113324..113412	mptgsb	Ser-tRNA gene (AGC); MYCPN
078	113856..115265	R02_orf469	MG061 homolog, MYCGE
079	115471..117165	R02_orf564o	hexosephosphate transport protein (uhpT); SALTY
080	118116..117217	D09_orf299	hypothetical protein (ywdF) homolog; BACSU
081	118123..118566	D09_orf147	hypothetical protein (A43259) homolog; ENTHR
082	118373..119539	D09_orf388	phosphoribosylpyrophosphate synthetase (prs); SYNP

Table 4. *Continued*

083	119518..120054	D09_orf178	hypothetical protein (yabF) homolog; BACSU
084	120036..120866	D09_orf276	hypothetical protein (yabC) homolog; BACSU
085	120853..121236	D09_orf127a	-
086	121404..121781	D09_orf125	MG055 homolog, MYCGE
087	121789..122751	D09_orf320	transcription antitermination factor (nusG); BACSU
088	124383..122719	D09_orf554	phosphomannomutase (cpsG); MYCPI
089	124774..124373	D09_orf133	cytidine deaminase (cdd); MYCPI
090	126050..124785	D09_orf421	thymidine phosphorylase (deoA); MYCPI
091	126711..126037	D09_orf224	deoxyribose-phosphate aldolase (deoC); MYCPN
092	127431..126715	D09_orf238	purine-nucleoside phosphorylase (deoD); ECOLI
093	127487..128839	D09_orf450	signal recognition particle protein (ffh); MYCMY
094	130278..129127	D09_orf383	S-adenosylmethionine synthetase 2 (metX); ECOLI
095	131221..130262	D09_orf319	o-sialoglycoprotein endopeptidase (gcp); PASHA
096	132678..131221	D09_orf485	putative lipoprotein, MG045 homolog, MYCGE
097	133523..132663	D09_orf286a	spermidine/putrescine transport system permease (potI); ECOLI
098	134376..133516	D09_orf286b	spermidine/putrescine transport system permease (potB); HAEIN
099	136060..134378	D09_orf560L	spermidine/putrescine transport ATP-binding prot (potA); ECOLI
100	137837..137466	D09_orf123	putative lipoprotein
101	139642..139376	D09_orf88	phosphocarrier protein HPr (ptsH); MYCCA
102	141633..139660	D09_orf657	putative lipoprotein, MG040 homolog, MYCGE
103	141816..142970	D09_orf384	aerobic glycerol-3-phospate dehydrogenase (glpD); ECOLI
104	142961..144487	D09_orf508	glycerol kinase (glpK), HAEIN
105	146845..144947	D09_orf632	MG288 homolog, MYCGE
106	148578..147022	D09_orf518	MG096 homolog, MYCGE
107	150522..149167	D09_orf451	pre-B cell enhancing factor homolog (pbeF); HUMAN
108	152171..150498	D09_orf557	aspartyl-tRNA synthetase (aspS); THEAQ
109	153387..152143	B01_orf414o	histidyl-tRNA synthetase (hisS); STREQ
110	153414..153989	B01_orf191	thymidine kinase (tdk); BACSU
111	154830..154036	B01_orf264	glycerol uptake facilitator (glpF); BACSU
112	157172..155154	B01_orf672	MG032 homolog, MYCGE
113	157794..157234	B01_orf186L	MG032 homolog, MYCGE
114	158048..158359	B01_orf103b	-
115	159270..158254	B01_orf338	MG032 homolog, MYCGE
116	159672..160020	B01_orf116L	-
	160267..160532	REPMP1	repetitive DNA sequence REPMP1
117	160694..160251	B01_orf147	-
118	162883..160862	B01_orf673	MG032 homolog, MYCGE
119	165055..163055	B01_orf666	MG032 homolog, MYCGE
120	165333..169664	B01_orf1443	DNA polymerase III (dnaE) alpha chain (3'-5' exonuclease); BACSU
121	169788..170324	B01_orf178	uracil phosphoribosyltransferase (upp); STRSL
122	170328..170654	B01_orf108	hypothetical protein (gi: 606093) homolog; ECOLI
123	171489..170878	B01_orf203	MG028 homolog, MYCGE
124	171995..171489	B01_orf168	MG027 homolog, MYCGE
125	172485..171913	B01_orf190	elongation factor P (efp) homolog; HAEIN
126	173405..172506	B01_orf299V	TrsB protein; YEREN
127	173438..174262	B01_orf274	-
128	175353..174265	B01_orf362	hypothetical protein (yyaF) homolog; BACSU
129	176220..175354	B01_orf288	fructose-bisphosphate aldolase (tsr); BACSU
130	176660..176220	B01_orf146	DNA-directed RNA polymerase delta subunit (rpoE); BACSU
131	178219..176681	B01_orf512	methionyl-tRNA synthetase (metS); BACST
132	179148..178219	B01_orf309	proline iminopeptidase (pip); NEIGO
133	180304..179132	D12_orf390o	heat shock protein DnaJ; BACSU
134	183442..180350	D12_orf1030	hypothetical helicase (yb95) homolog; YEAST
135	185356..183452	D12_orf634	transport ATP-binding protein (msbA); HAEIN
136	187139..185268	D12_orf623	transport ATP-binding protein (pmd1); SCHPO
	187233..187390	mptgi	Ile-tRNA(ATC), Ala-tRNA(GCA) genes; MYCPN
137	187475..188284	D12_orf269	5,10-methylene-tetrahydrofolate dehydrogenase (mtd1); HAEIN
138	188259..189125	D12_orf288	ribosomal protein S6 modification protein (rimK); ECOLI
139	189125..189982	D12_orf285	MG011 homolog, MYCGE
140	190597..189959	D12_orf212	DNA primase motif (dnaG); CLOAB
141	191472..190699	D12_orf257	-
142	192199..192906	D12_orf235	-
143	192931..193626	D12_orf231	putative lipoprotein
144	194207..193812	D12_orf131	-
145	195189..194404	D12_orf261	hypothetical protein (yabD) homolog; BACSU
146	196517..195189	D12_orf442	possible thiophene and furan oxidation protein (tdhF); BACSU
147	197280..196519	D12_orf253	DNA polymerase III subunit delta' (holB); ECOLI
148	197885..197253	D12_orf210	thymidylate kinase (CDC8) homolog, MYCGE
149	199152..197890	D12_orf420	seryl-tRNA synthetase (serS); BACSU
150	201643..199124	K05_orf839o	DNA gyrase subunit A (gyrA); STAAU
151	203595..201643	K05_orf650	DNA gyrase subunit B (gyrB); MYCPN
152	204626..203697	K05_orf309	DnaJ homolog protein; YEAST
153	205772..204630	K05_orf380	DNA polymerase III beta subunit (dnaN); STAAU
154	206520..207332	K05_orf270	protein (soj) homolog; BACSU
155	207319..208071	K05_orf250	-
156	208071..209390	K05_orf439	chromosomal replication initiator protein (dnaA); MYCCA
157	209458..210312	K05_orf284	sulfate transport ATP-binding protein (cysA); SYNP
158	210318..215966	K05_orf1882	-
159	215968..216987	K05_orf339	protein (devA) homolog; ANASP
160	217010..217156	K05_orf48	ribosomal protein L34 (rpL34); PROMI
161	217146..217502	K05_orf118V	RNaseP C5 chain (rnpA); MYCCA
162	217483..218640	K05_orf385	hypothetical protein 1 (S42122); MYCCA
163	218633..219424	K05_orf263V	S-adenosylmethionine-6-N',N'-adenosyl(rRNA) dimethyltransferase (ksgA); ECOLI
164	219411..220865	K05_orf484	glutamyl-tRNA synthetase (gltX); BACST
165	220846..222123	K05_orf425	MG461 homolog, MYCGE
166	223000..222680	K05_orf106	-
167	223391..223696	K05_orf101a	-
168	225039..224101	K05_orf312	L-lactate dehydrogenase (ldh); MYCHY
169	225210..225719	K05_orf169	hypothetical protein (HI0671) homolog; HAEIN
170	225719..226246	K05_orf175	hypoxanthine-guanine phosphoribosyltransferase (hpt); LACLA
171	226427..228556	K05_orf709	cell division protein (ftsH); BACSU
172	229109..230146	K05_orf345	MG456 homolog, MYCGE
173	231385..230186	K05_orf399	tyrosyl tRNA synthetase (tyrS); BACCA
174	231411..231833	K05_orf140	osmotically inducible protein (osmC); ECOLI
175	232705..231830	K05_orf291	UDP-glucose pyrophosphorylase (gtaB); BACSU
176	233448..232693	K05_orf251	MG452 homolog, MYCGE
177	233533..234717	K05_orf394	elongation factor TU (tuf); MYCGE
178	234876..235589	K05_orf237	homolog (degV) protein; BACSU
179	235596..236300	K05_orf234	MG449 homolog, MYCGE
180	236264..236719	K05_orf151	pilB homolog (fragment); HAEIN
181	236870..238369	K05_orf499	MG447 homolog, MYCGE
182	238451..238717	K05_orf88	ribosomal protein S16 (BS17); BACSU
183	238783..239415	K05_orf210	tRNA (guanine-N1)-methyltransferase (trmD); HUMAN
184	239399..239758	K05_orf119	ribosomal protein L19 (rpL19); BACST

Table 4. *Continued*

185	239774..240979	K05_orf401	hypothetical protein (P27712); SPICI
186	240948..241763	K05_orf271	MG442 homolog, MYCGE
187	242850..242236	E09_orf204o	protein P30, MYCPN
188	243127..243516	E09_orf129	putative lipoprotein
189	244320..243889	E09_orf143V	PTS system mannitol-specific component IIA (EIIA-MTL)(mtlF); STRMU
190	245395..244301	E09_orf364	mannitol-1-phosphate 5-dehyrogenase (EC 1.1.1.17)(mtlD); STRMU
191	246521..245382	E09_orf379	PTS system mannitol-specific component IIA (EIIA-MTL)(mtlA); STACA
192	247519..247824	E09_orf101	putative lipoprotein
193	247809..248219	E09_orf136L	-
194	249106..249516	E09_orf136	MG441 homolog, MYCGE
195	249627..250499	E09_orf290	putative lipoprotein, MG439 homolog, MYCGE
196	250522..251355	E09_orf277	putative lipoprotein, MG440 homolog, MYCGE
197	251355..252206	E09_orf283a	putative lipoprotein, MG439 homolog, MYCGE
198	252209..253060	E09_orf283b	putative lipoprotein, MG439 homolog, MYCGE
199	252981..253889	E09_orf302	MG440 homolog, MYCGE
200	253889..254782	E09_orf279	putative lipoprotein, MG439 homolog, MYCGE
201	254731..255561	E09_orf276	putative lipoprotein, MG440 homolog, MYCGE
202	255561..256463	E09_orf300	putative lipoprotein, MG439 homolog, MYCGE
203	256471..257334	E09_orf287o	MG439 homolog, MYCGE
204	258458..257331	E30_orf375	MG438 homolog, MYCGE
205	259665..258478	E30_orf395	CDP-diglyceride synthetase (cdsA); HAEIN
206	260219..259665	E30_orf184	ribosome releasing factor (frr); HAEIN
207	261354..260296	E30_orf352	-
208	262455..261910	C12_orf181o	-
209	263280..262537	C12_orf247	-
210	264090..263383	C12_orf235	uridylate kinase (pyrH); ECOLI
211	264988..264092	C12_orf298	elongation factor Ts (tsf); SPICI
212	265075..266289	C12_orf404	hypothetical protein (yfiB) homolog; SPICI
213	266342..267076	C12_orf244	triosephosphate isomerase (tim); ECOLI
214	267069..268595	C12_orf508	phosphoglycerate mutase (pgm); BACSU
215	268600..270318	C12_orf572	PEP-dependent HPr protein kinase phosphoryltransferase (Enzyme I) (ptsI); STRSL
216	270833..270315	C12_orf172	MG428 homolog, MYCGE
217	271393..270968	C12_orf141	MG427 homolog, MYCGE
218	271634..271437	C12_orf65	ribosomal protein L28 (rpL28); BACSU
219	273008..271656	C12_orf450	ATP-dependent RNA helicase (deaD); HAEIN
220	273166..273426	C12_orf86	ribosomal protein S15 (BS18); BACST
221	273431..275116	C12_orf561	MG423 homolog, MYCGE
222	275162..590313	C12_orf839	MG422 homolog, MYCGE
223	277659..280505	C12_orf948L	excinuclease ABC subunit A (uvrA); ECOLI
224	280514..282559	C12_orf681	DNA polymerase III subunit gamma and tau (dnaX); ECOLI
225	282590..283030	C12_orf146	ribosomal protein L13 (rpL13); ECOLI
226	283036..283434	C12_orf132	ribosomal protein S9 (rpS9); BACST
227	283864..284613	C12_orf249	restriction-modification enzyme subunit S1B (hsdS); MYCPU
228	284699..285703	C12_orf334	MG413 homolog, MYCGE
229	285639..286673	C12_orf344	MG415 homolog, MYCGE
230	286788..289781	C12_orf997	MG414 homolog, MYCGE
231	290023..291180	C12_orf385	MG412 homolog, MYCGE
232	291180..293135	C12_orf651V	phosphate transport system permease protein (pstA); ECOLI
233	293120..294109	C12_orf329	phosphate transport ATP-binding protein (pstB); ECOLI
234	294112..294789	C12_orf225	phosphate transport system regulatory protein (phoU); ECOLI
235	295259..294786	C12_orf157	peptide methionine sulfoxide reductase (pmsR), ECOLI
236	295314..296684	C12_orf456	enolase (eno) (EC 4.2.1.11); PLAFA
238	297129..298010	C12_orf293o	ATP synthase A chain (atpB); MYCGA
237	297163..296690	C12_orf157L	ATP synthase protein I (atpI); MYCGA
239	298013..298330	D02_orf105	ATP synthase C chain (atpE); MYCGA
240	298333..298956	D02_orf207	ATP synthase B chain (atpF); MYCGA
241	298949..299485	D02_orf178	ATP synthase delta chain (atpH); MYCGA
242	299488..301044	D02_orf518	ATP synthase alpha chain (atpA); MYCGA
243	301044..301883	D02_orf279	ATP synthase gamma chain (atpG); MYCGA
244	301883..303310	D02_orf475	ATP synthase beta chain (atpD); MYCGA
245	303313..303714	D02_orf133a	ATP synthase epsilon chain (atpC); MYCGA
246	303714..305423	D02_orf569	MG397 homolog, MYCGE
247	305423..305881	D02_orf152	galactose-6-phosphate isomerase subunit (lacA); STRMU
248	305799..306167	D02_orf122a	-
249	306393..306761	D02_orf122b	-
250	306862..308427	D02_orf521	putative lipoprotein, MG395 homolog, MYCGE
251	308950..310011	D02_orf353V	MG068 homolog, MYCGE
252	310168..310821	D02_orf217L	putative lipoprotein, MG395 homolog, MYCGE
253	310962..311435	D02_orf157L	MG395 homolog, MYCGE
254	311648..313243	D02_orf531	putative lipoprotein, MG395 homolog, MYCGE
255	313301..313753	D02_orf150	MG068 homolog, MYCGE
256	313629..314672	D02_orf347	MG067 homolog, MYCGE
257	314746..315654	D02_orf302	putative lipoprotein, MG068 homolog, MYCGE
258	315716..316123	D02_orf135L	MG067 homolog, MYCGE
259	316627..317304	D02_orf225L	MG068 homolog, MYCGE
260	317742..319061	D02_orf439	putative lipoprotein, MG068 homolog, MYCGE
261	319237..320034	D02_orf265V	MG068 homolog, MYCGE
262	320102..320524	D02_orf140	MG395 homolog, MYCGE
263	320666..320995	D02_orf109	-
264	321313..321011	D02_orf100	-
265	321751..322791	D02_orf346	MG068 homolog, MYCGE
266	322953..324173	D02_orf406	serine hydroxymethyltransferase (glyA); ACTAC
267	324608..324994	D02_orf128	-
268	325182..325532	D02_orf116	heat shock protein GroES; BACSU
269	325535..327166	D02_orf543	heat shock protein GroEL; BACSU
270	327180..328517	D02_orf445	nonspecified aminopeptidase; MYCSA
271	328621..330603	D02_orf660	lactococcin transport ATP-binding protein (lcnDR3); LACLA
272	330605..330994	D02_orf129	MG389 homolog, MYCGE
273	331116..331442	D02_orf108	MG388 homolog, MYCGE
274	331430..332305	D02_orf291	GTP-binding protein era homolog; STRMU
275	332405..335515	D02_orf1036o	protein P200; MYCPN
276	335519..336232	H03_orf237	glycerophosphoryl diester phosphodiesterase (glpQ); STAAU
277	336402..336860	H03_orf152	-
278	337074..338129	H03_orf351	NADP-dependent alcohol dehydrogenase (adh); THEBR
279	338333..339634	H03_orf433	GTP-binding protein (obg); BACSU
280	339627..340373	H03_orf248	probable NH(3)-dependent NAD(+) synthetase (outB); BACSU
281	341011..340370	H03_orf213	uridine kinase (udk); HAEIN
282	341065..342381	H03_orf438	arginine deiminase (arcA); PSEPU
	342382..342432	mptgab	Arg-tRNAgene (AGA); MYCPN
283	343166..342459	H03_orf235	MG381 homolog, MYCGE
284	343695..343120	H03_orf191	glucose inhibited division protein (gidB); ECOLI
285	345526..343688	H03_orf612	glucose inhibited division protein (gidA); ECOLI
286	345554..347167	H03_orf537	arginyl-tRNA synthetase (argS); BRELA
287	347210..347791	H03_orf193o	MG377 homolog (put. zinc protease), MYCGE

Table 4. *Continued*

288	347793..348107	G12_orf104	MG376 homolog, MYCGE
289	348107..349801	G12_orf564	threonyl-tRNA synthetase (thrSv); BACSU
290	349794..350603	G12_orf269	MG374 homolog, MYCGE
291	350610..351455	G12_orf281	MG373 homolog, MYCGE
292	351442..352605	G12_orf387	MG372 homolog, MYCGE
293	352598..353575	G12_orf325	hypothetical 28K protein (P1 operon) homolog; MYCPN
294	353562..354542	G12_orf326	hypothetical protein (HI0176) homolog; HAEIN
295	354597..356273	G12_orf558	MG369 homolog, MYCGE
296	356273..357259	G12_orf328a	fatty acid/phospholipid synthesis protein (plsX); ECOLI
297	357249..358097	G12_orf282a	ribonuclease III (rnc); ECOLI
298	360075..358081	G12_orf664	MG366 homolog, MYCGE
299	361010..360075	G12_orf311	methionyl-tRNA formyltransferase ((fmt); ECOLI
300	361671..361015	G12_orf218	MG364 homolog, MYCGE
301	361732..361995	G12_orf87	ribosomal protein S20 (rpsT); ECOLI
302	362178..362005	G12_orf57	ribosomal protein L32 (rpL32); HAEIN
303	362553..362185	G12_orf122	ribosomal protein L7/L12 ('A' type) (rpL7/L12); MICLU
304	363076..362591	G12_orf161	ribosomal protein L10 (rpL10); THEMA
305	363194..364432	G12_orf412	UV protection protein (mucB); ECOLI
306	365341..364418	G12_orf307	Holliday junction DNA helicase (ruvB); HAEIN
307	365936..365316	G12_orf206	Holliday junction DNA helicase (ruvA); ECOLI
308	366364..365942	G12_orf140b	-
309	366705..367877	G12_orf390	acetate kinase (ackA); BACSU
310	367885..368733	G12_orf282b	LicA protein (licA) homolog; HAEIN
311	368909..371056	G12_orf715	ATP-dependent protease binding subunit (clpB) homolog; HAEIN
312	371463..371053	G12_orf136	MG354 homolog, MYCGE
313	371612..371941	G12_orf109	MG353 homolog, MYCGE
314	373019..372465	G12_orf184	inorganic pyrophosphatase (ppa); THEAC
315	373074..373751	G12_orf225	-
316	374992..374006	G12_orf328b	MG350 homolog, MYCGE
317	376214..374973	G12_orf413	MG349 homolog, MYCGE
318	376807..377313	G12_orf168	-
	376824..377060	REPMP1	repetitive DNA sequence REPMP1
319	377903..378820	G12_orf305	putative lipoprotein, MG348 homolog, MYCGE
	378870..378945	mptgh	His-tRNA(CAC) gene; MYCPN
320	379607..378975	G12_orf210V	hypothetical protein (HI0340) homolog; HAEIN
321	380098..379598	G12_orf166b	hypothetical protein (ygl3) homolog; BACST
322	380141..382726	G12_orf861	isoleucine-tRNA ligase (ileS); STAAU
323	382844..383662	G12_orf272V	triacylglycerol lipase (lip) 3; MYCMY
324	383665..384711	G12_orf348	MG343 homolog, MYCGE
325	385804..386304	G12_orf166a	MG342 homolog, MYCGE
326	386397..390572	G12_orf1391o	RNA polymerase beta subunit (rpoB); BACSU
327	390576..394448	F04_orf1290	DNA-directed RNA polymerase beta' chain (rpoC); THEMA
328	394610..394972	F04_orf120	-
329	395489..395941	F04_orf150	-
330	396719..397183	F04_orf154	MG288 homolog, MYCGE
331	397214..397996	F04_orf260V	MG288 homolog, MYCGE
332	398608..399984	P02_orf458	MG096 homolog, MYCGE
333	401014..402297	P02_orf427	MG288 homolog, MYCGE
334	402844..404373	P02_orf509	MG288 homolog, MYCGE
335	405492..404401	P02_orf363V	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
336	407993..405612	P02_orf793	putative lipoprotein, MG260 homolog, MYCGE
337	408909..409670	P02_orf253	-
338	410118..409738	P02_orf126	-
339	411833..410688	P02_orf381	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
	412343..410580	REPMP5	repetitive DNA sequence REPMP5
340	413656..412388	P02_orf422V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	413701..412404	REPMP4	repetitive DNA sequence REPMP4
341	414691..414101	P02_orf196	-
	414718..414417	REPMP1	repetitive DNA sequence REPMP1
342	416640..415057	P02_orf527V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	416770..415161	REPMP2/3	repetitive DNA sequence REPMP2/3
343	417279..416788	P02_orf163	-
344	417961..417233	P02_orf242	L-ribulose-5-phosphate 4-epimerase (araD); ECOLI
345	418272..418703	P02_orf143	-
346	419131..421113	P02_orf660	hypothetical protein (yjfS) homolog; ECOLI
347	421405..421884	P02_orf159	hypothetical phosphotransferase protein (yjfU) homolog; ECOLI
348	421886..422542	P02_orf218	hypothetical protein (yjfV) homolog; ECOLI
349	422478..423395	P02_orf305	hypothetical protein (yjfW) homolog; ECOLI
350	424958..423534	P02_orf474	-
351	425032..426042	P02_orf336	recombination protein (recA); STAAU
352	426558..430460	P02_orf1300	putative lipoprotein, MG338 homolog, MYCGE
353	431060..430638	P02_orf140	MG337 homolog, MYCGE
354	432289..431063	P02_orf408	nitrogen fixation protein (nifS); HAEIN
356	432878..433828	P02_orf316	MG338 homolog, MYCGE
355	432936..432493	P02_orf147	-
	434119..434385	REPMP1	repetitive DNA sequence REPMP1
357	434245..434556	P02_orf103b	-
358	436086..435061	P01_orf341	hypothetical protein (yibD) homolog; ECOLI
359	436374..436955	P01_orf193	hypothetical protein (yihA) (era like) homolog; ECOLI
360	436939..439455	P01_orf838	valyl-tRNA synthetase (valS); BACST
361	439483..440076	P01_orf197	hypothetical protein (HI1366) homolog; HAEIN
362	440080..440787	P01_orf235	hypothetical protein (HI0315) homolog; HAEIN
363	440790..441419	P01_orf209	MG331 homolog, MYCGE
364	441446..442099	P01_orf217	cytidylate kinase (cmk); BACSU
365	442572..443450	P01_orf292	hypothetical protein (HI0136) (era like) homolog; HAEIN
366	443807..446908	P01_orf1033	MG328 homolog, MYCGE
367	446895..447701	P01_orf268	triacylglycerol lipase (lip) 2; MYCMY
368	447707..448588	P01_orf293	homolog (degV) protein; BACSU
369	448607..448768	P01_orf53	ribosomal protein L33 (rpL33); BACST
370	448768..449832	P01_orf354	X-Pro dipeptidase (pepX); LACDE
371	449873..450604	P01_orf243	-
	450647..451033	10saRNA	10saRNA; MYCGE
	451297..451058	rnpB RNA	RNaseP RNA; MYCGE
372	452076..451450	P01_orf208V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
373	452813..453118	P01_orf101	putative lipoprotein
374	453148..453570	P01_orf140	-
375	453614..454213	P01_orf199	-
	454252..453959	REPMP1	repetitive DNA sequence REPMP1
376	455967..454630	H08_orf445	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
377	456734..456261	H08_orf157a	-
	456769..454719	REPMP5	repetitive DNA sequence REPMP5
378	457621..456809	H08_orf270	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	457770..456825	REPMP4	repetitive DNA sequence REPMP4
379	458468..457773	H08_orf231	hypothetical protein (yzaC) homolog; BACSU

Table 4. *Continued*

380	458503..460200	H08_orf565	Na(+) translocating ATPase subunit J (ntpJ); ENTHR
	460165..460885	mptgv	Asn-tRNA(AAC), Glu-tRNA(GAA), Thr-tRNA(ACG), Val-tRNA(GTA), Thr-tRNA(ACA), Lys-tRNA(AAG), Leu-tRNA(CTA) genes; MYCPN
381	460960..462735	H08_orf591	MG321 homolog, MYCGE
382	462656..463129	H08_orf157b	MG321 homolog, MYCGE
383	463071..464060	H08_orf329V	adhesin P1 (group 2) homolog; MYCPN
384	464443..467460	H08_orf1005	putative lipoprotein, MG321 homolog, MYCGE
	467624..467717	mptgs	Ser-tRNA(TCC), Ser-tRNA(TCG) genes; MYCPN
385	467786..468649	H08_orf287	(cytochrome C oxidase polypeptide I (ctaD); BACSU)
386	468738..469319	H08_orf193	MG319 homolog, MYCGE
387	469340..470164	H08_orf274	30K adhesin-related protein; MYCPN
388	470178..472196	H08_orf672	cytadherence accessory protein (hmw3); MYCPN
389	472236..473345	H08_orf369	(competence locus E (comE3); BACSU)
390	473224..474168	H08_orf314	MG315 homolog, MYCGE
391	474180..475526	H08_orf448	MG314 homolog, MYCGE
392	475643..476434	H08_orf263	MG313 homolog, MYCGE
393	476498..479554	H08_orf1018	cytadherence accessory protein (hmw1); MYCPN
394	479577..480194	H08_orf205	ribosomal protein S4 (rpS4); BACSU
396	481119..485096	H08_orf1325	putative lipoprotein, MG309 homolog, MYCGE
395	481124..480255	H08_orf289	triacylglycerol lipase (lip) 3; Mycoplasma sp
397	485103..486332	H08_orf409	ATP-dependent RNA helicase (deaD); ECOLI
398	486317..486769	H08_orf150	putative lipoprotein, MG307 homolog, MYCGE
399	487390..487082	H08_orf102	-
400	487860..490040	H08_orf726	MG307 homolog, MYCGE
401	490196..490909	H08_orf237	putative lipoprotein, MG307 homolog, MYCGE
402	490965..492002	H08_orf345	MG307 homolog, MYCGE
403	492220..493938	H08_orf572o	MG307 homolog, MYCGE
404	494247..497981	A05_orf1244	putative lipoprotein, MG307 homolog, MYCGE
405	497991..499178	A05_orf395	MG306 homolog, MYCGE
406	499234..501021	A05_orf595	heat shock protein DnaK, ERYRH
407	501179..501991	A05_orf270L	abc transport ATP-binding protein (cbiO), SALTY
408	501886..503034	A05_orf382	abc transport ATP-binding protein (artP); ECOLI
409	503024..503977	A05_orf317	MG302 homolog, MYCGE
410	504008..505021	A05_orf337	glycerladehyde-3-phosphate dehydrogenase(gap), CLOPA
411	505024..506253	A05_orf409	phosphoglycerate kinase (pgk); THEMA
412	506291..507253	A05_orf320	phosphotransacetylase (pta); BACSU
413	508131..507259	A05_orf290	hypothetical protein (yidA) homolog; ECOLI
414	508316..511264	A05_orf982	P115 protein homolog (SGC3); MYCHR
415	511270..512316	A05_orf348	cell division protein (ftsY); ECOLI
416	512297..512605	A05_orf102	hypothetical 13.2 KD protein homolog (ylxM); BACSU
417	512605..512994	A05_orf129	MG296 homolog, MYCGE
418	512995..514107	A05_orf370	hypothetical protein (HI0174); HAEIN
419	514238..515665	A05_orf475	MG294 homolog (put. permease), MYCGE
420	515658..516383	A05_orf241a	glycerophosphoryl diester phosphodiesterase (glpQ); BACSU
421	516435..519137	A05_orf900	alanyl-tRNA synthetase (alaS); ECOLI
422	521188..519560	A05_orf542	transport system permease protein P69; MYCHR
423	521915..521181	A05_orf244	ATP-binding protein P29; MYCHR
424	523050..521908	A05_orf380V	high affinity transport system protein P37; MYCHR
425	524782..523301	A05_orf493	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
426	524892..525311	A05_orf139	-
	525343..523309	REPMP5	repetitive DNA sequence REPMP5
427	525388..526224	A05_orf278	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	526357..525404	REPMP4	repetitive DNA sequence REPMP4
428	526818..527576	A05_orf252	putative lipoprotein, MG440 homolog, MYCGE
	528050..527890	REPMP1	repetitive DNA sequence REPMP1
429	528164..527718	F11_orf148o	-
	528191..528045	REPMP1	repetitive DNA sequence REPMP1
430	530128..528527	F11_orf533L	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	530201..528684	REPMP2/3	repetitive DNA sequence REPMP2/3
431	532483..530201	F11_orf760	putative lipoprotein, MG260 homolog, MYCGE
432	532711..535350	F11_orf879	-
	535464..535390	mptgwa	Trp-tRNA (TGA) gene; MYCPN
433	535709..535455	F11_orf84	(acyl carrier protein; STRGA)
434	536337..535744	F11_orf197	MG286 homolog, MYCGE
435	537384..536344	F11_orf346	MG285 homolog, MYCGE
436	537733..537365	F11_orf122a	MG284 homolog, MYCGE
437	539329..537878	F11_orf483	putative prolyl-tRNA synthetase (proS); YEAST
438	539611..540093	F11_orf160	transcription elongation factor (greA); RICPR
	540123..540573	mptrna	Tyr-tRNA (TAC), Glu-tRNA (CAA), Lys-tRNA (AAA), Leu-tRNA (TTA), Gly-tRNA (GGA) genes; MYCPN
439	540861..542609	F11_orf582	MG281 homolog, MYCGE
440	542671..543534	F11_orf287	MG280 homolog, MYCGE
441	543534..544190	F11_orf218	MG279 homolog, MYCGE
442	546388..544187	F11_orf733	stringent response protein (spoT); ECOLI
443	546644..549307	F11_orf887	MG277 homolog, MYCGE
444	549474..549875	F11_orf133	adenine phosphoribosyltransferase (apt); HAEIN
445	549943..551382	F11_orf479	NADH oxidase (nox); ENTFA
446	551403..552479	F11_orf358a	pyruvate dehydrogenase E1-alpha subunit (pdhA); ACHLA
447	552501..553484	F11_orf327	pyruvate dehydrogenase E1-beta subunit (pdhB); ACHLA
448	553803..555011	F11_orf402	dihydrolipoamide acetyltransferase component (E2) (pdhC); ACHLA
449	555012..556385	F11_orf457	dihydrolipoamide dehydrogenase (pdhD); BACST
450	556412..557431	F11_orf339	lipoate protein ligase (lplA); ECOLI
451	557803..558879	F11_orf358b	MG269 homolog, MYCGE
	558904..558982	4.5s RNA	4.5S RNA; MYCPN
452	559027..559716	F11_orf229	hypothetical protein (yaaF) homolog; BACSU
453	559751..560095	F11_orf114	MG267 homolog, MYCGE
454	560096..562477	F11_orf793o	leucyl-tRNA synthetase (leuS); BACSU
455	562480..563328	A19_orf282	hypothetical protein (yidA) homolog; ECOLI
456	563860..563258	A19_orf200	hypothetical protein (HI0890) homolog; HAEIN
457	564732..563854	A19_orf292	hypothetical protein (yidA) homolog; ECOLI
458	565711..564878	A19_orf277	formamidopyrimidine-DNA glycosylase (fpg); BACFI
459	566586..565711	A19_orf291	DNA polymerase I (polA, 5'-3' exonuclease) homolog; STRPN
460	569208..566590	A19_orf872	DNA polymerase III alpha subunit (dnaE); HAEIN
	569524..569598	mptga	Arg-tRNA gene (CGA); MYCPN
461	569863..573285	A19_orf1140	-
462	573664..574053	A19_orf129	-
463	574399..575088	A19_orf229V	-
464	576117..576731	A19_orf204	-
465	578517..576742	A19_orf591	-
466	578671..579306	A19_orf211	-
	579725..578587	REPMP4	repetitive DNA sequence REPMP4
	581534..580008	REPMP2/3	repetitive DNA sequence REPMP2/3
467	581562..579349	A19_orf737V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
468	582203..582964	H91_orf253	putative lipoprotein
469	583638..583096	H91_orf180	-

Table 4. *Continued*

	583663..583392	REPMP1	repetitive DNA sequence REPMP1
470	585295..584327	H91_orf322	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
471	586044..585226	H91_orf272	hypothetical 130K protein homolog (orf6, P1 operon); MYCPN
	586110..584114	REPMP5	repetitive DNA sequence REPMP5
472	586934..586128	H91_orf268	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
473	589311..587278	H91_orf677	MG260 homolog, MYCGE
474	589658..589350	H91_orf102	putative lipoprotein, MG260 homolog, MYCGE
475	591151..589790	H91_orf453	possible protoporphyrinogen oxidase (hemK); ECOLI
476	592230..591151	H91_orf359V	peptide chain release factor 1 (RF1; prfA);BACSU
477	592524..592231	H91_orf97	ribosomal protein L31 (rpL31); ECOLI
478	593345..592569	H91_orf258	MG256 homolog, MYCGE
	593426..593353	mptgg	Trp-tRNA(TGG) gene; MYCPN
479	595179..593575	H91_orf534	MG255 homolog, MYCGE
	595211..595283	mptgg	Gly-tRNA(GGC) gene; MYCPN
480	595347..597323	H91_orf658	DNA ligase (lig); ECOLI
481	597304..598617	H91_orf437	cysteinyl-tRNA synthetase (cysS); BACSU
482	598620..599348	H91_orf242a	hypothetical protein (yacO) (rRNA methylase) homolog; BACSU
483	599370..600719	H91_orf449	glycyl-tRNA synthetase (grs1); YEAST
484	600703..602565	H91_orf620	DNA primase (dnaG); BACSU
485	602618..604117	H91_orf499	RNA polymerase sigma-A factor (sigA); BACSU
486	604101..604742	H91_orf213	MG248 homolog; MYCGE
487	604748..605467	H91_orf239	hypothetical protein (ygiH) homolog; ECOLI
488	606304..605459	H91_orf281	MG246 homolog, MYCGE
489	606788..606294	H91_orf164	5-formyl tetrahydrofolate cyclo-ligase (HI0858) homolog; HAEIN
490	608873..607743	H91_orf376	Type 1 restriction enzyme (hsdR) homolog; ECOLI
491	609427..609080	H91_orf115	-
492	610177..609557	H91_orf206	Type 1 restriction enzyme (hsdR) homolog; ECOLI
493	611772..611122	H91_orf216	-
494	612987..611995	H91_orf330	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
495	614997..613366	H91_orf543	type I restriction enzyme (hsdM); ECOLI
496	617285..615138	H91_orf715	DNA helicase II (mutB1); HAEIN
497	618937..617348	H91_orf529	DNA helicase (pcrA) homolog; STAAU
498	619615..618941	H91_orf224	MG243 homolog, MYCGE
499	621513..619615	F10_orf632o	MG242 homolog, MYCGE
500	623381..621516	F10_orf621	MG241 homolog, MYCGE
501	623625..624500	F10_orf291	MG240 homolog, MYCGE
502	626726..624501	F10_orf741	-
503	627693..626713	F10_orf326	protein (bcrA) homolog; BACLI
504	629948..627696	F10_orf750	putative ABC transport permease
505	632530..630143	F10_orf795	ATP-dependent protease (lon); BACSU
506	633935..632601	F10_orf444	trigger factor (tig); HAEIN
507	634844..633960	F10_orf294	MG237 homolog, MYCGE
508	635310..634834	F10_orf158	MG236 homolog, MYCGE
509	636124..635264	F10_orf286	endonuclease IV (nfo); ECOLI
510	636431..636117	F10_orf104	ribosomal protein L27 (rpL27); BACSU
511	636726..636424	F10_orf100a	hypothetical protein (ysxB) homolog; BACSU
512	637021..636719	F10_orf100b	ribosomal protein L21 (rpL21); BACSU
513	639333..637168	F10_orf721	ribonucleoside-diphosphate reductase (nrdE); SALTY
514	639818..639357	F10_orf153	MG230 homolog, MYCGE
515	640840..639821	F10_orf339	ribonucleotide reductase 2 (nrdF); SALTY
516	641329..640847	F10_orf160	dihydrofolate reductase (EC 1.5.1.3)(dhfr); LACLA
517	642317..641331	F10_orf328	thymidylate synthase (thyA); STAAU
518	644200..642689	F10_orf503	general amino acid permease GAP1 homolog; YEAST
519	645650..644175	F10_orf491	hypothetical protein (gi: 710640) homolog (put. amino acid permease); CLOPE
520	646835..645693	F10_orf380	cell division protein (ftsZ); BACSU
521	648100..646841	F10_orf419	MG223 homolog, MYCGE
522	649029..648103	F10_orf308	hypothetical protein (yabC) homolog; ECOLI
523	649444..649019	F10_orf141b	hypothetical protein (yabB) homolog; ECOLI
	649775..649699	mptgac	Arg-tRNA gene (CGC); MYCPN
524	649845..650117	F10_orf90	MG220 homolog, MYCGE
525	650856..650200	F10_orf218	-
526	651919..650846	F10_orf357	-
527	657390..651934	F10_orf1818	cytadherence accessory protein (hmw2); MYCPN
528	658627..657410	F10_orf405	protein P65; MYCPN
529	660458..658761	F10_orf565	-
530	661390..660461	F10_orf309	carbamate kinase (EC 2.7.2.2) (arcC); PSEAE
531	662214..661393	H10_orf273o	ornithine carbamoyl transferase (otc1); ECOLI
532	663058..662462	H10_orf198	arginine deiminase (arcA); MYCCA
533	663675..662959	H10_orf238	arginine deiminase (arcA); MYCCA
	664617..663872	mptgc	Cys-tRNA(TGC), Pro-tRNA(CCA), Met-tRNA(ATG), Ile-tRNA(ATG), Ser-tRNA(TCA), fMet-tRNA(ATG), Asp-tRNA(GAC) and Phe-tRNA(TTC) genes; MYCPN
534	666181..664655	H10_orf508	pyruvate kinase (pyk); LACLA
535	667173..666187	H10_orf328	6-phosphofructokinase (pfk); ECOLI
536	667819..667193	H10_orf208	hypothetical protein (P35155) homolog; BACSU
537	669323..667803	H10_orf506	dihyrofolate reductase (dyr) homolog protein; ENTFC
538	670124..669324	H10_orf266	1-acyl-sn-glycerol-3-phosphate acyltransferase (plsB); YEAST
539	670471..670112	H10_orf119	-
540	670923..670474	H10_orf149	MG211 homolog, MYCGE
541	671792..671130	H10_orf220L	-
542	672461..671841	H10_orf206	-
543	672500..673054	H10_orf184	prolipoprotein signal peptidase (lsp); STACA
544	673054..673983	H10_orf309	hypothetical protein (yceC) homolog; ECOLI
545	673967..674557	H10_orf196	MG208 homolog, MYCGE
546	674987..674550	H10_orf145L	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
547	675689..675126	H10_orf187V	HsdS1B protein homolog; MYCPU
548	678142..675779	A65_orf787o	putative lipoprotein, MG260 homolog, MYCGE
549	679094..678738	A65_orf118	-
	680988..679736	REPMP2/3	repetitive DNA sequence REPMP2/3
550	681222..679825	A65_orf465V	adhesin P1 (group 2) homolog; MYCPN
551	682245..681325	A65_orf306	protein (prrB) homolog, ECOLI
552	685088..682704	A65_orf794	putative lipoprotein, MG260 homolog, MYCGE
	686360..686126	REPMP1	repetitive DNA sequence REPMP1
553	686379..686032	A65_orf115	-
554	688090..687590	A65_orf166	MG260 homolog, MYCGE
555	689578..688445	A65_orf377	MG260 homolog, MYCGE
556	691498..689789	A65_orf569	MG139 homolog, MYCGE
557	693374..691629	A65_orf581	GTP-binding membrane protein (lepA); HAEIN
558	694573..693374	A65_orf399V	YefE protein homolog; ECOLI
559	696002..694533	A65_orf489	lysyl-tRNA synthetase (lysS); BACSU
560	696047..696904	A65_orf285	MG135 homolog, MYCGE
561	697178..696876	A65_orf100	hypothetical protein (yaaK) homolog; BACSU
562	697200..698000	A65_orf266	MG133 homolog, MYCGE
563	697969..698403	A65_orf144	hypothetical protein (hit1) homolog; YEAST
564	701122..700367	A65_orf251a	putative lipoprotein, MG440 homolog, MYCGE

Table 4. *Continued*

565	703155..701674	A65_orf493	hypothetical protein (ysr1) homolog; MYCMY
566	703498..703145	A65_orf117	MG129 homolog, MYCGE
567	704277..703498	A65_orf259	hypothetical protein (HI0072) homolog; HAEIN
568	704714..704277	A65_orf145	hypothetical protein (ygl1) homolog; STRVR
569	704771..705811	A65_orf346	tryptophanyl-tRNA synthetase (trpS); HAEIN
570	706664..705819	A65_orf281	hypothetical protein (gi: 973220) homolog; ECOLI
571	706984..706676	A65_orf102	thioredoxin (trx); YEAST
572	708477..707050	A65_orf475	MG123 homolog, MYCGE
573	710602..708467	A65_orf711	DNA topoisomerase I (topA); BACSU
574	711574..710639	A65_orf311	high affinity ribose transport protein (rbsC); HAEIN
575	713127..711574	A65_orf517	MG120 homolog, MYCGE
576	714862..713144	A65_orf572	hypothetical ABC transporter (yjcW) homolog; ECOLI'
577	715893..714877	A65_orf338	UDP-glucose 4-epimerase (galE); STRTR
578	716545..715874	A65_orf223	MG117 homolog, MYCGE
579	717293..716538	A65_orf251b	MG116 homolog, MYCGE
580	718497..717814	A65_orf227	phosphatidylglycerophosphate synthase (pgsA); HAEIN
581	719821..718454	K04_orf455o	asparaginyl-tRNA synthetase (asnS); ECOLI
582	720475..719828	K04_orf215L	D-ribulose-5-phosphate 3 epimerase (cfxE); ALCEU
583	721745..720453	K04_orf430	phosphoglucose isomerase B (pgiB); BACST
584	722603..721767	K04_orf278L	hypothetical protein (yjeQ) homolog; ECOLI
585	723759..722590	K04_orf389	probable protein serine/threonine kinase (YKT3); CAEEL
586	724529..723750	K04_orf259	protein phoshatase 2C homolog (ptc1); YEAST
588	725070..725720	K04_orf216	polypeptide deformylase (def); HAEIN
587	725248..724529	K04_orf239	5'guanylate kinase (gmk); HAEIN
589	726297..725689	K04_orf202	MG105 homolog, MYCGE
590	728477..726297	K04_orf726	virulence associated protein homolog (vacB); HAEIN
591	729593..728751	K04_orf280	MG103 homolog, MYCGE
592	730530..729583	K04_orf315	thioredoxin reductase (trxB); EUBAC
593	731191..730523	K04_orf222	MG101 homolog, MYCGE
594	732602..731166	G07_orf478o	protein (pet112) homolog; YEAST
595	734028..732592	G07_orf478V	amidase homolog (S47454); YEAST
596	735470..734031	G07_orf479	MG098 homolog, MYCGE
597	736390..735668	G07_orf240	uracil DNA glycosylase (ung); ECOLI
598	737668..736415	G07_orf417	MG288 homolog, MYCGE
599	739760..738396	G07_orf454	putative lipoprotein, MG095 homolog, MYCGE
600	741185..739764	G07_orf473	replicative DNA helicase (dnaC); BACSU
601	741621..741172	G07_orf149	ribosomal protein L9 (rpL9); BACST
602	741938..741624	G07_orf104b	ribosomal protein S18 (rpS18); ECOLI
603	742428..741928	G07_orf166	single-stranded DNA binding protein (ssb); HAEIN
604	743075..742428	G07_orf215	ribosomal protein S6 (rpS6); ECOLI
605	745198..743132	G07_orf688	elongation factor G (fus); THEAQ
606	745688..745221	G07_orf155	ribosomal protein S7 (rpS7); BACST
607	746161..745742	G07_orf139	ribosomal protein S12 (rpS12); BACST
608	747359..746190	G07_orf389b	prolipoprotein diacylglyceryl transferase (lgt); ECOLI
609	748287..747349	G07_orf312	MG085 homolog, MYCGE
610	749157..748288	G07_orf289	hypothetical protein (yacA) homolog; BACSU
611	749716..749150	G07_orf188	peptidyl-tRNA hydrolase homolog (pth); HAEIN
612	750396..749716	G07_orf226	ribosomal protein L1 (rpL1); BACST
613	750809..750396	G07_orf137	ribosomal protein L11 (RPL11); THEMA
614	753420..750865	G07_orf851	oligopeptide transport ATP-binding protein (oppF); BACSU
615	754654..753383	G07_orf423	oligopeptide transport ATP-binding protein (oppD); BACSU
616	755786..754656	G07_orf376	oligopeptide transport system permease protein (amiD); STRPN
617	756948..755779	G07_orf389a	oligopeptide transport system permease protein (oppB); BACSU
618	757224..757640	G07_orf138	MG076 homolog, MYCGE
619	760729..757637	G07_orf1030	protein P100; MYCPN
620	761241..760834	G07_orf135	MG074 homolog, MYCGE
621	763217..761244	G07_orf657	excinuclease ABC subunit B (uvrB); ECOLI
622	765618..763192	G07_orf808	preprotein translocase (secA); BACSU
623	768223..765605	G07_orf872V	MG(2+) transport ATPase, P-typ 1 (mgtA); ECOLI
624	769100..768216	G07_orf294	ribosomal protein S2 (rpS2); SPIPL
625	772532..769710	GT9_orf940o	PTS system, glucose-specific IIABC component (EIIABC-GLC); BACSU
626	772584..772925	GT9_orf113	-
627	774296..772980	GT9_orf438V	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	774345..773095	REPMP4	repetitive DNA sequence REPMP4
628	775203..774757	GT9_orf148	MG260 homolog, MYCGE
	775230..774929	REPMP1	repetitive DNA sequence REPMP1
629	775949..775566	GT9_orf127	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
630	776809..775868	GT9_orf313	ADP1_MYCPN adhesin P1 precursor homolog; MYCPN
	777250..775724	REPMP2/3	repetitive DNA sequence REPMP2/3
631	778005..777289	GT9_orf238	type I restriction enzyme ecokI specificity protein (hsdS) homolog; HAEIN
632	780875..778479	GT9_orf798	putative lipoprotein, MG260 homolog, MYCGE
633	783441..781159	GT9_orf760	putative lipoprotein, MG185 homolog, MYCGE
634	784494..783535	GT9_orf319V	adenine-specific methyltransferase EcoRI (mte1); ECOLI
635	786329..784494	GT9_orf611	oligoendopeptidase F (pepF); LACLA
636	787053..786322	GT9_orf243V	pseudouridylate synthase I (hisT); ECOLI
637	788350..787046	GT9_orf434	MG181 homolog, MYCGE
638	789254..788343	GT9_orf303	histidine transport ATP-binding protein (hisP); ECOLI
639	790066..789242	GT9_orf274	sulfate transport ATP-binding protein (cysA); SYNP
640	790424..790050	GT9_orf124a	ribosomal protein L17 (rpL17); BACSU
641	791410..790427	GT9_orf327	RNA polymerase alpha core subunit (rpoA); BACSU
642	791781..791416	GT9_orf121	ribosomal protein S11 (rpS11); BACST
643	792155..791781	GT9_orf124b	ribosomal protein S13 (rpS13); BACSU
644	792268..792155	GT9_orf37	ribosomal protein L36 (rpL36); CHLTR
645	792515..792279	GT9_orf78	initiation factor 1 (infA); BACSU
646	793261..792515	GT9_orf248	methionine amino peptidase (map); BACSU
647	793908..793261	GT9_orf215	adenylate kinase (adk); BACST
648	795335..793902	GT9_orf477	preprotein translocase subunit (secY); MYCCA
649	795790..795335	GT9_orf151	ribosomal protein L15 (rpL15); MYCCA
650	796453..795794	GT9_orf219	ribosomal protein S5 (rpS5); BACSU
651	796807..796457	GT9_orf116b	ribosomal protein L18 (rpL18); BACST
652	797362..796808	GT9_orf184	ribosomal protein L6 (rpL6); MYCCA
653	797797..797369	GT9_orf142	ribosomal protein S8 (rpS8); MYCCA
654	797976..797791	GT9_orf61	ribosomal protein S14 (rpS14); MYCCA
655	798520..797978	GT9_orf180b	ribosomal protein L5 (rpL5); HAEIN
656	798858..798523	GT9_orf111a	ribosomal protein L24 (rpL24); BACST
657	799226..798858	GT9_orf122	ribosomal protein L14 (rpL14); BACST
658	799487..799230	GT9_orf85	ribosomal protein S17 (rpS17); MYCCA
659	799822..799487	GT9_orf111b	ribosomal protein L29 (rpL29); THEMA
660	800241..799822	VXpSPT7_orf139o	ribosomal protein L16 (rpL16); MYCCA
661	801062..800241	VXpSPT7_orf273	ribosomal protein S3 (rpS3); MYCCA
662	801618..801064	VXpSPT7_orf184	ribosomal protein L22 (rpL22); HAEIN
663	801808..801545	VXpSPT7_orf87	ribosomal protein S19 (rpS19); MYCBO
664	802671..801808	VXpSPT7_orf287a	ribosomal protein L2 (rpL2); MYCCA
665	803384..802671	VXpSPT7_orf237	ribosomal protein L23 (rpL23); THEMA

Table 4. *Continued*

666	804025..803387	VXpSPT7_orf212	ribosomal protein L4 (rpL4); MYCCA
667	804888..804025	VXpSPT7_orf287b	ribosomal protein L3 (rpL3); MYCCA
668	805228..804902	VXpSPT7_orf108	ribosomal protein S10 (rpS10); THEMA
669	805660..805322	VXpSPT7_orf112	
670	806869..805907	VXpSPT7_orf320	putative lipoprotein, MG149 homolog, MYCGE
671	808328..806991	VXpSPT7_orf445	MG148 homolog, MYCGE
672	809615..808482	VXpSPT7_orf377	MG147 homolog, MYCGE
673	810876..809602	VXpSPT7_orf424	hemolysin (hlyC) homolog protein; HAEIN
674	811711..810902	VXpSPT7_orf269	hypothetical protein (yaaC) homolog; PSEFL
675	812932..811724	VXpSPT7_orf402	MG144 homolog, MYCGE
676	813298..812948	VXpSPT7_orf116	ribosome binding factor A homolog (rbfA); ECOLI
677	815154..813301	VXpSPT7_orf617	protein synthesis initiation factor 2 (infB); BACST

noteworthy: the lack of the ribosomal protein S1, of the peptide chain release factor 2 (RF2) and of the glutaminyl-tRNA synthetase. So far, quite a number of Gram-positive bacteria including *Bacillus* or *Lactobacillus* species also lack the S1 protein and the glutaminyl-tRNA synthetase (46).

One of the functions of the S1 protein is to bind the mRNA to the 30S small ribosomal subunit. Therefore, it was argued that ribosomal binding sites in front of many genes (47) of *B.subtilis* compensate for the missing S1 protein. The Shine–Dalgarno sequences are so well conserved, that they could be used routinely as a good indicator for proposing ORFs in the *B.subtilis* genome sequencing projects, but this does not apply to *M.pneumoniae*. The Shine–Dalgarno sequence is in many instances not well conserved or missing altogether, even in genes for which we know the translational initiation sites from independent studies.

Of the 20 standard tRNA-synthetases, the glutaminyl-tRNA synthetase is the only one not detected in *M.pneumoniae*. Studies on tRNA synthetases in Gram-positive bacteria have indicated that this enzyme is dispensable. *Bacillus subtilis* solves this problem by charging the $tRNA^{Gln}$ first with glutamate which is subsequently converted to glutamine by an amido transferase. The glutamyl tRNA synthetase aminoacylates both $tRNA^{Glu}$ and $tRNA^{Gln}$. The corresponding amido transferase has not yet been identified in *M.pneumoniae*, therefore it is still an open question as to how glutamine is bound to its tRNA.

Finally, the modified codon usage by *M.pneumoniae*, reading UGA as tryptophan instead of a stop codon, requires the absence of the peptide chain release factor 2 (RF2) and the presence of the release factor 1 (RF1). The latter recognizes the stop codons UAG and UAA and RF2 the stop codons UGA and UAA. Since the UGA codon is frequently located within a gene it is essential to exclude RF2 to prevent the premature termination of proteins.

Surface structure, cytadherence-associated proteins and cell division

This category comprises the adhesins and the cytadherence associated proteins, including the components of the cytoskeleton-like structure, the function of which is probably to stabilize and maintain the shape of the wall-less mycoplasma, to direct proteins to certain regions in the membrane and to keep them in these positions (2). Adherence to the receptor(s) of the host cell depends on the tip structure. The correct assembly of the adhesin P1 (E07_orf1627) and the 30 kDa adhesin-related protein on the tip structure (H08_orf274) is necessary for attachment. The tip structure is an interesting example for bacterial cellular asymmetry (48).

The cytadherence-associated proteins were originally defined by hemadsorption-negative mutants which had lost certain proteins like the so called high molecular weight proteins HMW1, HMW2 and HMW3, the adhesin P1 and the proteins named A, B and C (2,28). B and C are most probably the gene products of the ORF6 gene of the P1 operon (40 kDa protein = C, 90 kDa protein = B). The gene for A is still unknown. Another criterion for a putative protein of the cytoskeleton-like structure is its partitioning into the Triton X-100 insoluble fraction after treating *M.pneumoniae* with this detergent. This fraction is ill defined and comprises ~50 proteins, of which only a subfraction is associated with the cytoskeleton and/or cytadherence. The following proteins have been identified as most likely components of a cytoskeleton (2): HMW1 (H08_orf1018), HMW2 (F10_orf1818; Krause, submitted), HMW3 (H08_orf672), P200 (D02_orf1036o) (49), P65 (F10_orf405) (27). These proteins, with the exception of HMW2, share some common peculiar features, like an extended acidic proline rich domain and an abnormal migration in SDS-PAGE (49). The adhesin P1 is mainly distributed in the membrane fraction and to a lesser extent in the Triton X-100 insoluble fraction (50).

A large number of proposed ORFs contain sequences with high similarities to subregions of either the P1 protein or the ORF6 gene product of the P1 operon. The coding DNA sequences correspond to the repetitive DNA sequences RepMP2/3 (P1), RepMP4 (P1) and RepMP5 (ORF6). Preliminary experiments indicate that the proposed ORFs are not expressed under standard laboratory conditions. It has been observed that another independent isolate of *M.pneumoniae*, the strain FH, carries a different copy of RepMP2/3, RepMP4 and RepMP5 in its P1 operon than the *M.pneumoniae* strain M129 which is the subject of this paper (51,52). All experimental data so far show that only the repetitive sequences which are part of the P1 operon are expressed. The exchange of these copies presumably takes place by gene conversion as was indicated by DNA sequence analysis of the corresponding RepMP5 sequences in *M.pneumoniae* strains M129 and FH. Different is the situation with RepMP1, copies of which seem to be part of several expressed proteins. RepMP1-specific antibodies recognize several proteins on western blots of *M.pneumoniae* protein extracts (26).

Only little is known about cell division in *M.pneumoniae*. The lack of mutants, especially of conditional mutants, has prevented a detailed analysis. So far, the two proteins FtsZ and FtsH are classified as cell division proteins in analogy to their function in other bacteria (53). Other genes involved in chromosome partitioning or septum formation have not been identified in *M.pneumoniae*. Interesting problems to study might include the possible interaction of FtsZ with components of the cytoskeleton-like structure, which seems to play a key role in cell division, or the effects of cellular asymmetry on cell division and the formation of daughter cells. Other genes known to be involved in cell division in *E.coli*, the muk and min genes or additional fts genes were not found in *M.pneumoniae* (53).

Lipoproteins

Altogether 46 proteins were identified as lipoproteins based on the following characteristic lipoprotein-specific features (54): (i) one or more basic amino acids among the first 5–7 amino acids of the N-terminus, (ii) a hydrophobic signal peptide and (iii) a cysteine residue immediately downstream of the signal peptide, which is available for modification by the transfer of the diacylglyceryl moiety from glycerophospholipid to its sulfhydryl group. The precursor prolipoprotein with the modified cysteine is subsequently cleaved in *M.pneumoniae* by a specific signal peptidase (signal peptidase II). The modified cysteine will then be the first amino

acid of the processed protein. The cleavage site including the cysteine and the three (positions −3, −2 and −1) upstream located amino acids, is to some extent conserved (−3: 37×L, 6×F, 1×A, 1×V; −2: 19×S, 10×A, 8×T, 6×V, 2×I; −1: 37×A, 7×S, 1×G).

The number of lipoproteins in *M.pneumoniae* is relatively high compared with the Gram-negative bacteria *E.coli* and *H.influenzae*. Even in the closely related *M.genitalium* only 21 putative lipoproteins could be found by analyses of the published data (9).

The lipoproteins of *M.pneumoniae* can be divided into six subgroups based on sequence similarities; also included in these groups are proteins with similarities to lipoproteins but without the lipoprotein signature at the N-terminal end. Quite a number of these proposed genes with high similarities are organized in tandem. For instance seven lipoproteins and one protein without the lipobox but with otherwise extended similarities are located between genome positions 249 627 and 256 463 (cosmid pcosMPE09). A gene family, with 13 proposed ORFs including five lipoproteins, is located between 306 862 and 320 524 (cosmid pcosMPD02). Presently it is unclear whether all of the proposed genes are expressed.

In vivo labelling of *M.pneumoniae* with ^{14}C-labelled palmitic acid and protein analysis by SDS–PAGE reveal, instead of the expected 46 lipoproteins, only between 20 and 25 lipoproteins (Pyrowolakis, unpublished data). This discrepancy could be explained either by a regulated expression which only allows some of the several tandemly organized lipoproteins to be synthesized or that the labelling with palmitic acid was not sensitive enough or that some lipoproteins carry fatty acids other than palmitic acid. Only four of all the proposed lipoproteins show significant similarities to other bacterial genes beside the ones from *M.genitalium.* These include A05_orf380V [high affinity transport system P37 with unknown specificity from *Mycoplasma hyorhinis* (55)], D09_orf384 (aerobic glycerol-3-phosphate dehydrogenase, glpD), H03_orf213 (uridine kinase) and D02_orf207 (ATP synthase b subunit (atpF).

The processing of the prolipoprotein to the mature lipoprotein in *E.coli* requires the three enzymes prolipoprotein diacylglyceryl transferase, prolipoprotein signal peptidase and apolipoprotein transacylase. We find in *M.pneumoniae* only the transferase which catalyzes the thioether linkage between the diacylglycerol and the cysteine and the peptidase which cleaves in front of the cysteine following the signal peptide. The transacylase could not be identified either in *M.pneumoniae* nor in *M.genitalium* (9). Therefore it is still an open question if a third fatty acid is linked to the cysteine by an amide bond as has been found for lipoproteins of *E.coli.*

The absence of a periplasmic space provides reasons for the existence of a large number of lipoproteins. For surface-exposed proteins which have to function on the outside, anchoring them via long chain fatty acids at the *M.pneumoniae* cell membrane is an efficient way. Already known examples are substrate-binding proteins of transport systems or proteins possibly involved in antigenic variation for evasion of the immune system of the host, as has been shown for other mycoplasmas (56). Nothing is known about the fate of the cleaved signal peptides, as to whether they are degraded or recycled.

Transport systems

In light of the scarcity of metabolic pathways and the marked dependence on exogenous nutrients (Table 1, Fig. 5), we expected *M.pneumoniae* to code for many transport systems to compensate for its inability to synthesize essential compounds like amino acids. Three different transport systems, mainly involved in import, were found in *M.pneumoniae*: (i) the ABC transporter system (57) consisting of two ATP-binding, two membrane-spanning and one substrate-binding domain which are frequently present on separate polypeptides, but sometimes also consist of two or three different domains located on the same peptide (D12_orf634 or D12_orf623), (ii) the phosphoenolpyruvate: carbohydrate phosphotransferase system (PTS), (58) and (iii) facilitated diffusion systems with transmembrane proteins functioning as specific carriers. *Mycoplasma pneumoniae* codes for 43 genes involved in the above mentioned transport systems according to the present status of annotation. In addition, there are several proposed proteins with 6 or 12 transmembrane segments which are candidates for membrane-spanning domains of transport systems. The relatively low number of proteins listed in Table 1 indicates that at least some of the systems might not be very substrate specific, e.g. the transport systems for amino acids. Transport systems for histidine, glutamine, an ORF showing significant similarity to a probable aromatic amino acid permease from yeast and an ABC transport system for oligopeptides were identified based on similarity of the ATP-binding domains of ABC transporters.

Surprisingly, we could not identify a transport system for the precursors for RNA and DNA synthesis, namely adenine, guanine, uracil and thymine which are essential components of mycoplasma growth media.

In this context one has to be aware of the ambiguity in the identification of ABC transport proteins on the basis of sequence similarity of the ATP-binding proteins with respect to the predicted substrate to be transported, since database searches indicate numerous candidates with different specificities but with very similar, high score values. All the annotations in this paper were done on the basis of the highest score values. Therefore it might be possible that the predicted specificity disagrees with the *in vivo* activity in *M.pneumoniae.* Additional information from similarities to transmembrane domains or the substrate-binding proteins is only rarely at hand, since, in general, similarities among these domains are not well conserved. Even in positive examples, the score values are relatively low. Sometimes additional circumstantial evidence is derived from an operon-like organisation of the genes coding for ABC transporters, e.g. the unspecified ABC transporter consisting of the proteins P69, P29 and P37 from nucleotide 519 560 to 523 050 (A05_orf542, A05_orf244 and A05_orf380V). A05_orf542 could act as the membrane-spanning domain, A05_orf244 as the ATP-binding domain and A05_orf380V, as a putative lipoprotein which could function as a substrate-binding protein. These proteins were also identified by their significant similarity to the corresponding genes in *M.hyorhinis* (55).

In *M.pneumoniae* the ABC transport system for oligopeptides consists of two different transmembrane [G07_orf376 = amiD (= oppC in *B.subtilis*); G07_orf389a = oppB] and ATP-binding domains (G07_orf851 = oppF , G07_orf423 = oppD). It is also organized in an operon-like arrangement from nucleotide 750 865 to 756 948. In striking contrast to *B.subtilis*, the substrate-binding domain (oppA) is absent in *M.pneumoniae*. Since an oppA homolog is also absent in *M.genitalium* a sequencing or annotation error seems unlikely. It remains to be experimentally determined whether the substrate-binding protein is dispensable or is part of one of the transmembrane or ATP-binding proteins.

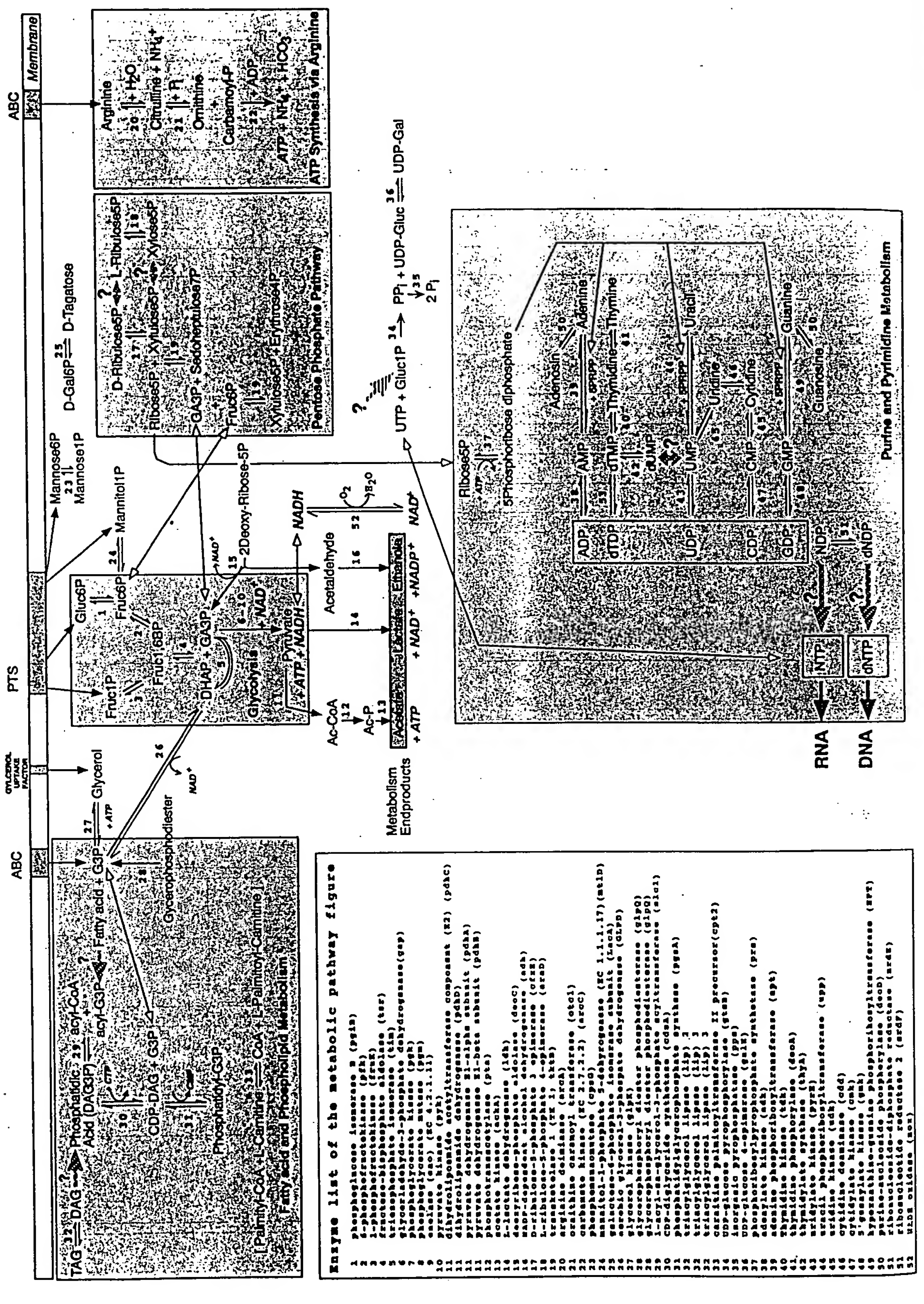


Figure 5. Schematic diagram of the metabolic pathways of *M.pneumoniae* deduced from Table 1. Shaded arrows with question marks indicate missing enzymatic activities.

It is also possible that one or more of the lipoproteins function as substrate-binding proteins.

There is also evidence for bacterial ABC export systems in *M.pneumoniae* (59). For example D12_orf634 (msbA), D12_orf623 (pmd1) and D02_orf660 (lcnDR3) have the conserved ATP binding motif and the membrane-spanning domains on the same polypeptide. In addition D12_orf623 and D12_orf634 show also significant similarities to multidrug resistance proteins of different organisms.

Among the proposed PTS transport systems, we identified one for glucose and one for mannitol. They are similar to the homologous systems from several Gram-positive bacteria, with a EIIA and EIIBC domains on two separate polypeptides for the mannitol transport system and with three domains (EIIABC) of enzyme II in one polypeptide for the glucose transport system.

Besides glucose and mannitol, fructose also seems to be imported by the PTS system. According to our data the fructose-permease II component R02_orf694 (fruA) contains all three domains of enzyme II in one gene (EIIABC). In addition, R02_orf694 and the 1-phosphofructokinase (fruK, R02_orf300) are probably in one operon, but we do not find fruF which is also part of the fructose operon in enteric bacteria (58).

Protein secretion

Both, Gram-positive and Gram-negative bacteria have a well conserved protein translocation system. The components identified which are part of the well characterized *E.coli* system (60) include cytosolic chaperones or regulators [trigger factor, SecB, DnaK, SRP (a ribonucleoprotein composed of 4.5 S RNA and Ffh) and FtsY] which deliver the protein to a membrane receptor (SecA). The receptor is also supposed to function as a motor, pushing the protein across the membrane via specific protein channels (SecY, SecG, SecE, SecD and SecF). The secreted proteins to be transported carry an N-terminal signal peptide which will be removed by a signal peptidase (SPaseI). Two routes of export have been proposed either via SecB and SecA or by SRP. The protein secretion system in *M.pneumoniae* is less complex (Table 1). So far, the trigger factor, DnaK, SRP, FtsY and SecA have been identified. From the channel-forming proteins only SecY is present but SecG, SecF, SecE, SecD and the cytosolic receptor protein SecB are missing. Also absent is the signal peptidase SPaseI although computer-assisted motif prediction programs indicate the presence of corresponding substrates (signal peptides). The simplified protein export system might be a reflection of the fact that *M.pneumoniae* is only surrounded by a cytoplasmic membrane. Another problem concerns refolding of secreted proteins which are normally exported in an unfolded stage. Refolding might be catalyzed by chaperones which have to function on the cell surface (60). This might impose a special problem on the wall-less bacteria in general, since they do not possess a periplasmic space which could prevent proteins from diffusing. To anchor the proposed chaperones on the cell surface as lipoproteins would be a possible way to solve this problem.

Nucleotide synthesis: purine and pyrimidine salvage pathways

Guanine, guanosine, uracil, thymine, thymidine, cytidine, adenine and adenosine may serve as precursors for nucleic acids and nucleotide coenzymes, as determined in nutritional studies of *Mollicutes*. These components can be used for the synthesis of ribonucleotides by the salvage pathway as predicted from the enzymes listed (Table 1, Fig. 5). The ribonucleotides are converted to deoxyribonucleotides by ribonucleoside–diphosphate reductase, an enzyme complex formed by the gene products of nrdE (F10_orf721) and nrdF (F10_orf339). Adenine, guanine and uracil can be metabolized directly to the corresponding nucleoside monophosphates by the enzymes adenine phosophoribosyltransferase (apt, F11_orf133), hypoxanthine-guanine phosphoribosyltransferase (hpt, K05_orf175) and uracil phosphoribosyltransferase (upp, B01_orf178). Uridylate, adenylate and guanylate kinases catalyze the generation of ADP, GDP and UDP. Surprisingly, we could not find the nucleoside diphosphate kinase (ndk), the key enzyme for the conversion from NDP to NTP. This finding is in agreement with data from the genomic sequence analysis of *M.genitalium*.

Another important enzyme, the CTP synthetase which converts UTP to CTP is also missing. Therefore the only route for the synthesis of CTP appears to be from cytidine to CMP by uridine kinase (H03_orf213) and to CDP by cytidylate kinase (P01_orf217). Deoxythymidine monophosphate (dTMP) could be either synthesized by thymidine kinase (tdk, B01_orf191) or by thymidylate synthase (thA, F10_orf328).

It will be of special interest to experimentally identify the enzyme(s) of *M.pneumoniae* which convert NDPs to NTPs, since such an enzymatic activity seems to be essential.

Carbohydrate metabolism and energy conservation

The ability to metabolize glucose and/or arginine and use it for the ATP synthesis is one of the key features in classification of *Mollicutes*. *Mycoplasma pneumoniae* is listed in Bergey's manual of systematic bacteriology as a glucose fermenter but not as an arginine-hydrolyzing species (61). This contrasts with our sequencing results, since the three enzymes involved in the arginine degradation pathway, arginine deiminase (H03_orf438), ornithine carbamoyltransferase (H10_orf273) and carbamate kinase (F10_orf309) are present according to our sequence data. The arginine deiminase gene occurs twice but one copy is inactive due to a raster-mutation resulting in two proposed ORFs (H10_orf198 and H10_orf238) corresponding to the N-terminal and C-terminal halves of a complete deiminase. The change in reading frame was also confirmed by sequencing of directly amplified genomic DNA. All these proposed ORFs are organized in an operon-like arrangement except for the deiminase (H03_orf438) which seems to be expressed as a single gene located far away from the mentioned operon. Included in this operon is a proposed protein (F10_orf565) with 12 predicted transmembrane domains indicative of a putative permease.

Glucose, fructose and mannitol are transported by the PTS system into the cell and further degraded by the Embden–Meyerhof–Parnas (EMP) pathway to pyruvate. All enzymes required for this pathway have been identified. The second pathway for metabolizing glucose, the pentose phosphate pathway, is incomplete in *M.pneumoniae*. We found only the enzymes ribulose-5-phosphate-3-epimerase and transketolase (Fig. 5). Glucose-6-phosphate dehydrogenase (G6Pde), 6-phospho-gluconate dehydrogenase (6PGde), and a transaldolase are missing. These data agree with enzymatic studies showing that G6Pde and 6PGde are absent in mycoplasmas (62).

Pyruvate can be further metabolized by two alternative reactions, either to lactate by lactate dehydrogenase (K05_orf312) or to acetyl-CoA by the pyruvate dehydrogenase complex and further to acetate by the phosphotransacetylase (A05_orf320, pta) and the acetate kinase (G12_orf390, ackA). The pyruvate dehydrogenase complex consists of E1α (F11_orf358a) E1β (F11_orf327), the two subunits of the pyruvate dehydrogenase, the dihydrolipoamide acetyltransferase E2 (F11_orf402) and the dihydrolipoamide dehydrogenase E3 (F11_orf457). The corresponding genes are clustered (nt 549 943–557 431; pcosMPF11); part of this cluster also contains the genes coding for NADH oxidase (nox, F11_orf479) and lipoate protein ligase (lplA, F11orf339). The later enzyme joins lipoic acid in an amide linkage to the ε amino group of a lysine residue of the dihydrolipoamide acetyltransferase.

Membrane phospho- and glycolipid synthesis

In *M.pneumoniae* strain FH the following membrane phospho- and glycolipids have been found: digalactosyldiacylglycerol, trigalactosyldiacylglycerol, glucosylgalactosyldiacylglycerol, phosphatidylglycerol (PG) and diphosphatidylglycerol (DPG) (63). Since *M.pneumoniae* FH and *M.pneumoniae* M129 are very similar we assume that both strains carry essentially the same genes for phospho- and glycolipid-synthesis.

About 10 genes are required for the synthesis of the above-mentioned lipids; but according to our DNA sequence analysis only three of the expected genes could be unambiguously identified. They code (Fig. 5) for the enzymes 1-acylglycerol-3-phosphate acyltransferase (plsC; gene name in *Saccharomyces cerevisiae* is slc1), phosphatidic acid cytidyltransferase (cdsA) and glycerolphosphate phosphatidyltransferase (pgsA). These enzymes are involved in the biochemical pathway for the synthesis of PG and DPG. Missing are the glycerol-3-phosphate acyltransferase (plsB) catalysing the synthesis of 1-acylglycerol-3-phosphate (acyl-G3P) from glycerol-3-phosphate (G3P), the phosphatidylglycerol phosphate phosphatase which converts phosphatidylglycerol-3-phosphate to PG and finally the cardiolipin synthetase (cls) which synthesizes DPG from PG. Interestingly, we find a gene homologous to the plsX gene from *E.coli* which is involved in membrane lipid synthesis in an undefined manner. The glycolipid synthesis could start with phosphatidic acid and would probably require a phosphatidic acid phosphatase and several UDP-glucosyl- or galactosyltransferases. None of these enzymes could be identified by similarity searches in databases.

As expected from biochemical studies no gene involved in fatty acid or cholesterol synthesis was determined in the sequence analysis. These components are incorporated as such from the medium.

An interesting enzyme is the proposed carnitine palmitoyl-transferase encoded by C09_orf600, which might be involved in the modifacation of exogenous phosphatidylcholine (67).

CONCLUSIONS

It is impossible to address each proposed *M.pneumoniae* gene in this paper. We have tried to cover the most important categories of functions and point to genes which should be present, but could not be found by our applied methods. Typical examples are the missing diphosphonucleoside kinase for the conversion of (d)NDPs to (d)NTPs, and the substrate binding domain (oppA) for the oligopeptide ABC transporter. In addition, we could not find any indication for a number of genes/proteins, which should be there based on experimental evidence. *Mycoplasma pneumoniae* has been shown to be motile and to exhibit chemotactic behaviour (64). Motility genes are difficult to identify since the motility of *M.pneumoniae* is independent of pili or flagella and it is not yet known which are potential candidates. Therefore, any progress in this field depends on the isolation of mutants. Furthermore, none of the components of the chemotactic signal pathway, the Che proteins, which are well conserved among bacteria, or any other 'two-component signal transduction system' could be detected. Chemotactic behaviour in *M.pneumoniae* is difficult to study. While it might be possible that these bacteria are chemotaxis negative, only additional experiments will clarify this point.

It has been reported that *M.pneumoniae* produces hydrogen peroxide considered to be a pathogenicity factor (17). Therefore, to protect itself from oxidative stress one would expect to find the standard enzymes dealing with these stress factors like catalase, superoxide dismutase or peroxidase, but we have no similarity based evidence that these enzymes exist in *M.pneumoniae*. Experimental data on this topic are also inconsistent (62).

The results of our sequence analysis explain quite well the kind of changes which have led to the observed reduction of the genome size in *M.pneumoniae* from the presumed genome size of several million base pairs of the ancestral bacteria. The main cause is the loss of complete anabolic (no amino acid synthesis) and metabolic pathways and of genes for the synthesis of complex structures like the bacterial cell wall which requires a large number of genes. In addition, for several processes like DNA repair, DNA recombination, cell division or protein secretion, the number of genes involved is smaller than in the more complex bacteria.

No significant changes were observed in the size of individual genes which resemble more or less their counterparts in *E.coli* or *B.subtilis*. The occasionally observed smaller intergenic regions, like those found in the ATPase operon, do not appear to significantly contribute to the overall genome size reduction.

In contrast with the loss of complete pathways we frequently observed the amplification of complete genes or segments of genes (see sections on lipoprotein families or on the repetitive DNA sequences RepMP2/3, RepMP4 and RepMP5). In these two instances the obvious advantage would be the potential of expressing antigenic variants of surface-exposed proteins.

The various truncated genes which are also present in full length copies e.g. arginine deiminase (H03_orf438 and H03_orf238), DNA primase (H91_orf620 and D12_orf212) and the dihydrofolate reductase (H10_orf506 and F10_orf160) might be relics of recombination events which took place in the course of the process of evolution.

Finally among the many proposed proteins are a few which share the highest similarity over their entire length with a eukaryotic protein. The most prominent examples are the pre-B cell enhancing factor (pbeF, D09_orf451) and the carnitine palmitoyltransferase II precursor (cpt2, C09_orf600). Both might be candidates for examples of horizontal gene transfer, but at the present state of analysis a definitive answer cannot be given.

It will be the main task of future studies to reconcile the experimental evidence and the DNA sequence-based predictions, i.e. to indentify the genes for observed functions and vice versa, and to assign functions to proposed open reading frames with hitherto unknown functions.

One obvious topic is the comparative analysis between the completely sequenced genomes of the closely related species *M.pneumoniae* and *M.genitalium* (9). Since the present paper is already very voluminous we decided to publish this analysis in an additional paper (Himmelreich *et al.*, in preparation).

ACKNOWLEDGEMENTS

We thank R. Frank and A. Bosserhoff for the synthesis of oligonucleotides, B. Reiner for her expertise in computer data analysis, Raphael Mosbach for his technical assistance concerning hardware problems, U. Leibfried for technical assistance, I. Schmidt for preparing the manuscript, D. Hofmann and H. Göhlmann for reading of the manuscript and H. Schaller for financial assistance and his encouragement throughout our work. We thank S. Razin, A. Wieslander, K. Dybvig, K. Sitaraman, R. Walker, H. Neimark and R. Miles who read drafts of this publication. Their corrections, critical comments and suggestions helped us very much. This research was supported by a grant from the Deutsche Forschungsgemeinschaft (He 780/5-1–He 780/5-4) and by the Fonds der Chemischen Industrie.

REFERENCES

1 Chanock, R. M., Dienes, L., Eaton, M. D., Edward, D. G., Freundt, E. A., Hayflick, L., Hers, J. F. P., Jensen, K. E., Liu, C., Marmion, B. P., Morton, H. E., Mufson, M. A., Smith, P. F., Somerson, N. L. and Taylor-Robinson, D. (1963) *Science*, 140, 662.
2 Krause, D. C. (1996) *Mol. Microbiol.*, **20**, 247–253.
3 Jacobs, E. (1991) *Rev. Med. Microbiol.*, **2**, 83–90.
4 Dybvig, K. (1990) *Annu. Rev. Microbiol.*, **44**, 81–104.
5 Morowitz, H. J. (1984) *Isr. J. Med. Sci.*, **20**, 750–753.
6 Razin, S. (1992) *FEMS Microbiol Lett*, **100**, 423–431.
7 Bove, J. M. (1993) *Clin. Infect. Dis.*, **17 Suppl 1**, 10–31
8 Peterson, S. N., Hu, P. C., Bott, K. F. and Hutchison, C. A. d. (1993) *J. Bacteriol.*, **175**, 7918–7930
9 Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M. *et al.* (1995) *Science*, **270**, 397–403.
10 Hilbert, H., Himmelreich, R., Plagens, H. and Herrmann, R. (1996) *Nucleic Acids Res.*, **24**, 628–639.
11 Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W. and Gillevet, P. M. (1995) *Mol. Microbiol.*, **16**, 955–967
12 Sterky, F., Holmberg, A. and Uhlen, M. (1996) HUGO'96, Heidelberg, Germany.
13 Glass, J. L., Glass, J. S., Lefkowitz, E. J., Chen, E. Y. and Cassel, G. H. (1996) IOM Letters, USA, Vol. 4, pp. 12., Proc. Meet. Int. Org. Mycoplasm., Orlando, Florida.
14 Inamine, J. M., Loechel, S. and Hu, P. C. (1988) *Gene*, **73**, 175–183.
15 Wenzel, R. and Herrmann, R. (1989) *Nucleic Acids Res.*, **17**, 7029–7043.
16 Su, C. J., Chavoya, A. and Baseman, J. B. (1988) *Infect Immunol.*, **56**, 3157–3161.
17 Almagor, M., Yatziv, S. and Kahane, I. (1983) *Infect. Immunol.*, **41**, 251–256.
18 Sanger, F., Nicklen, R. and Coulson, A. R. (1977) *Proc. Natl Acad. Sci. USA*, **79**, 5463–5467.
19 Bairoch, A. and Boeckmann, B. (1991) *Nucleic Acids Res.*, **19**, 2247–2249.
20 Barker, W. C., George, D. G., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.
21 Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
22 Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) *J. Mol. Biol.*, **215**, 403–410.
23 Bairoch, A. (1992) *Nucleic Acids Res.*, **20**, 2013–2018.
24 Inamine, J. M., Ho, K. C., Loechel, S. and Hu, P. C. (1990) *J. Bacteriol.*, **172**, 504–506.
25 Nakai, K. and Kanehisa, M. (1991) *Proteins: Struct., Funct. Genet.*, **11**, 95–110.
26 Proft, T. and Herrmann, R. (1994) *Mol. Microbiol.*, **13**, 337–348.
27 Proft, T., Hilbert, H., Layh Schmitt, G. and Herrmann, R. (1995) *J. Bacteriol.*, **177**, 3370–3378.
28 Razin, S. and Jacobs, E. (1992b) *J. Gen. Microbiol.*, **138**, 407–422.
29 Ruland, K., Wenzel, R. and Herrmann, R. (1990) *Nucleic Acids Res.*, **18**, 6311–6317.
30 Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. *et al.* (1995) *Science*, **269**, 496–512.
31 Riley, M. (1993) *Microbiol. Rev.*, **57**, 862–952.
32 Baker, T. A. and Wickner, S. H. (1992) *Annu. Rev. Genet.*, **26**, 447–477.
33 Mills, L. B., Stanbridge, E. J., Sedwick, W. D. and Korn, D. (1977) *J. Bacteriol.*, **132**, 641–649.
34 Barnes, M. H., Tarantino, P. M., Jr., Spacciapoli, P., Brown, N. C., Yu, H. and Dybvig, K. (1994) *Mol. Microbiol.*, **13**, 843–54
35 Koonin, E. V. and Bork, P. (1996) *Trends Biochem. Sci.*, **21**, 128–129.
36 Camerini-Otero, R. D. and Hsieh, P. (1995) *Annu. Rev. Genet.*, **29**, 509–552.
37 Demple, B. and Harrison, L. (1994) *Annu. Rev. Biochem.*, **63**, 915–948.
38 Sancar, A. and Sancar, G. B. (1988) *Annu. Rev. Biochem.*, **57**, 29–67.
39 Haldenwang, W. G. (1995) *Microbiol. Rev.*, **59**, 1–30
40 Hyman, H. C., Gafny, R., Glaser, G. and Razin, S. (1988) *J. Bacteriol.*, **170**, 3262–3268.
41 Moran, C. P. j., Lang, N., LeGrice, S. F. J., Lee, G., Stephens, M., Sonnenshein, A. L., Pero, J. and Losik, R. (1982) *Mol. Gen. Genet.*, **186**, 339–346.
42 Das, A. (1993) *Annu. Rev. Biochem.*, **62**, 893–930.
43 Hecker, M., Schumann, W. and Voelker, U. (1996) *Mol. Microbiol.*, **19**, 417–428.
44 Parkinson, J. S. (1993) *Cell*, **73**, 857–871.
45 Simoneau, P., Li, C. M., Loechel, S., Wenzel, R., Herrmann, R. and Hu, P. C. (1993) *Nucleic Acids Res*, **21**, 4967–4974.
46 Breton, R., Watson, D., Yaguchi, M. and Lapointe, J. (1990) *J. Biol. Chem.*, **265**, 18248–18255.
47 Shine, J. and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA*, **71**, 1342–1346.
48 Shapiro, L. (1993) *Cell*, **73**, 841–855.
49 Proft, T., Hilbert, H., Plagens, H. and Herrmann, R. (1996) *Gene*, **171**, 79–82.
50 Kahane, I., Tucker, S., Leith, D. K., Morrison, P. J. and Baseman, J. B. (1985) *Infect. Immunol.*, **50**, 944–946.
51 Su, C. J., Chavoya, A., Dallo, S. F. and Baseman, J. B. (1990) *Infect. Immunol.*, **58**, 2669–2674.
52 Ruland, K., Himmelreich, R. and Herrmann, R. (1994) *J. Bacteriol.*, **176**, 5202–5209
53 Vicente, M. and Errington, J. (1996) *Mol. Microbiol.*, **20**, 1–7.
54 Sankaran, K., Gupta, S. D. and Wu, H. C. (1995) *Methods Enzymol*, **250**, 683–697
55 Gilson, E., Alloing, G., Schmidt, T., Claverys, J. P., Dudler, R. and Hofnung, M. (1988) *EMBO J.*, **7**, 3971–3974.
56 Citti, C. and Wise, K. S. (1995) *Mol. Mircobiol.*, **18**, 649–660.
57 Higgins, C. F. (1992) *Annu. Rev. Cell Biol.*, **8**, 67–113.
58 Postma, P. W., Lengeler, J. W. and Jacobson, G. R. (1993) *Microbiol. Rev.*, **57**, 543–594
59 Fath, M. J. and Kolter, R. (1993) *Microbiol. Rev.*, **57**, 995–1017.
60 Schatz, G. and Dobberstein, B. (1996) *Science*, **271**, 1519–1526
61 Freundt, E. A. and Razin, S. (1984) In Krieg, N. R. and Holt, J. G. e. (eds), *Bergey's Manual of Systematic Bacteriology*, Vol. 1. Williams and Wilkins, Baltimore, pp. 742–770.
62 Pollack, J. D. (1992) In Maniloff, J., McElhaney, R. N., Finch, L. R. and Baseman, J. B. e. (eds), *Mycoplasmas—Molecular Biology and Pathogenesis*. American Society for Microbiology, Washington, DC, pp. 181–200.
63 Plackett, P., Marmion, B. P., Shaw, E. J. and Lemke, R. M. (1969) *Aust. J. Exp. Biol. Med. Sci.*, **47**, 171–195.
64 Kirchhoff, H. (1992) In Maniloff, J., McElhaney, R. N., Finch, L. R. and Baseman, J. B. e. (eds.), *Mycoplasmas—Molecular Biology and Pathogenesis*. Americam Society for Microbiology, Washington, DC, pp. 289–308.
65 Matic, I., Rayssiguier, C. and Radman M. (1995) *Cell*, **80**, 507–515.
66 Atkins, J. F. and Gesteland, R. F. (1996) *Nature*, **379**, 769–771
67 Rottem, S., Adar, L., Gross, Z., Ne'Eman, Z. and Davis, P. J. (1986) *J. Bacteriol.*, **167**, 299–304.

e complete genome sequence the Gram-positive bacterium *acillus subtilis*

unst[1], N. Ogasawara[2], I. Moszer[3], A. M. Albertini[4], G. Alloni[4], V. Azevedo[5], M. G. Bertero[3,4], P. Bessières[5], A. Bolotin[5], S. Borchert[6], orriss[7], L. Boursier[3], A. Brans[8], M. Braun[9], S. C. Brignell[10], S. Bron[11], S. Brouillet[3,12], C. V. Bruschi[13], B. Caldwell[14], V. Capuano[5], I. Carter[10], S.-K. Choi[15], J.-J. Codani[16], I. F. Connerton[17], N. J. Cummings[17], R. A. Daniel[18], F. Denizot[19], K. M. Devine[20], A. Düsterhöft[9], . Ehrlich[5], P. T. Emmerson[21], K. D. Entian[6], J. Errington[18], C. Fabret[19], E. Ferrari[14], D. Foulger[18], C. Fritz[9], M. Fujita[22], Y. Fujita[23], S. Fuma[24], alizzi[4], N. Galleron[5], S.-Y. Ghim[15], P. Glaser[3], A. Goffeau[25], E. J. Golightly[26], G. Grandi[27], G. Guiseppi[19], B. J. Guy[10], K. Haga[28], J. Haiech[19], . Harwood[10], A. Hénaut[29], H. Hilbert[9], S. Holsappel[11], S. Hosono[30], M.-F. Hullo[3], M. Itaya[31], L. Jones[32], B. Joris[8], D. Karamata[33], asahara[2], M. Klaerr-Blanchard[3], C. Klein[6], Y. Kobayashi[30], P. Koetter[6], G. Koningstein[34], S. Krogh[20], M. Kumano[24], K. Kurita[24], A. Lapidus[5], ardinois[8], J. Lauber[9], V. Lazarevic[33], S.-M. Lee[35], A. Levine[36], H. Liu[28], S. Masuda[30], C. Mauël[33], C. Médigue[3,12], N. Medina[36], . Mellado[37], M. Mizuno[30], D. Moestl[9], S. Nakai[2], M. Noback[11], D. Noone[20], M. O'Reilly[20], K. Ogawa[24], A. Ogiwara[38], B. Oudega[34], I. Park[15], V. Parro[37], T. M. Pohl[39], D. Portetelle[40], S. Porwollik[7], A. M. Prescott[18], E. Presecan[3], P. Pujic[5], B. Purnelle[25], G. Rapoport[1], Rey[26], S. Reynolds[33], M. Rieger[41], C. Rivolta[33], E. Rocha[3,12], B. Roche[36], M. Rose[6], Y. Sadaie[22], T. Sato[30], E. Scanlan[20], S. Schleich[3], Schroeter[7], F. Scoffone[4], J. Sekiguchi[42], A. Sekowska[3], S. J. Seror[36], P. Serror[5], B.-S. Shin[15], B. Soldo[33], A. Sorokin[5], E. Tacconi[4], akagi[43], H. Takahashi[28], K. Takemaru[30], M. Takeuchi[30], A. Tamakoshi[24], T. Tanaka[44], P. Terpstra[11], A. Tognoni[27], V. Tosato[13], S. Uchiyama[42], Vandenbol[40], F. Vannier[36], A. Vassarotti[45], A. Viari[12], R. Wambutt[46], E. Wedler[46], H. Wedler[46], T. Weitzenegger[39], P. Winters[14], A. Wipat[10], Yamamoto[42], K. Yamane[24], K. Yasumoto[28], K. Yata[22], K. Yoshida[23], H.-F. Yoshikawa[28], E. Zumstein[6], H. Yoshikawa[2] & A. Danchin[3]

itut Pasteur, Unité de Biochimie Microbienne, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France
a Institute of Science and Technology, Graduate School of Biological Sciences, Ikoma, Nara 630-01, Japan
itut Pasteur, Unité de Régulation de l'Expression Génétique, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France
artimento di Genetica e Microbiologia, Universita di Pavia, Via Abbiategrasso 207, 27100 Pavia, Italy
RA, Génétique Microbienne, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex, France
itut für Mikrobiologie, J. W. Goethe-Universität, Marie Curie Strasse 9, 60439 Frankfurt/Maine, Germany
itut für Genetik und Mikrobiologie, Humboldt Universität, Chausseestrasse 17, D-10115 Berlin, Germany
tre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6, Sart Tilman, B-4000 Liège, Belgium
AGEN GmbH, Max-Volmer-Strasse 4, D-40724 Hilden, Germany
partment of Microbiological, Immunological and Virological Sciences, The Medical School, University of Newcastle, Framlington Place, Newcastle upon Tyne NE2 4HH, UK
partment of Genetics, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands
elier de BioInformatique, Université Paris VI, 12 rue Cuvier, 75005 Paris, France
GEB, AREA Science Park, Padriciano 99, I-34012 Trieste, Italy
enencor International, 925 Page Mill Road, Palo Alto, California 94304-1013, USA
cterial Molecular Genetics Research Unit, Applied Microbiology Research Division, KRIBB, PO Box 115, Yusong, Taejon 305-600, Korea
RIA, Domaine de Voluceau, PB 105, 78153 Le Chesnay Cedex, France
stitute of Food Research, Department of Food Macromolecular Science, Reading Laboratory, Earley Gate, Whiteknights Road, Reading RG6 6BZ, UK
William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, UK
aboratoire de Chimie Bactérienne, CNRS BP 71, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 09, France
epartment of Genetics, Trinity College, Lincoln Place Gate, Dublin 2, Republic of Ireland
epartment of Biochemistry and Genetics, The Medical School, University of Newcastle, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK
adioisotope Center, National Insitute of Genetics, Mishima, Shizuoka-ken 411, Japan
epartment of Biotechnology, Faculty of Engineering, Fukuyama University, Higashimura-cho, Fukuyama-shi, Hiroshima 729-02, Japan
nstitute of Biological Sciences, Tsukuba University, Tsuiuba-shi, Ibaraki 305, Japan
aculté des Sciences Agronomiques, Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud, 2-20 B-1348 Louvain-la-Neuve, Belgium
ovo Nordisk Biotech, 1445 Drew Avenue, Davis, California 95616-4880, USA
niricerche, Via Maritano 26, San Donato Milanese, 20097 Milan, Italy
nstitute of Molecular and Cellular Biology, The University of Tokyo, Bunkyo-ku, Tokyo 113, Japan
aboratoire Génome et Informatique, Université de Versailles, Bâtiment Buffon, 45 Avenue des États-Unis, 78035 Versailles Cedex, France
aculty of Agriculture, Tokyo University of Agriculture and Technology, Fuchu, Tokyo 183, Japan
Mitsubishi Kasei Institute of Life Sciences, 11 Minamyiooa, Machida-shi, Tokyo 194, Japan
nstitut Pasteur, Service d'Informatique Scientifique, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France
nstitut de Génétique et Biologie Microbiennes, Université de Lausanne, 19 rue César Roux, 1005 Lausanne, Switzerland
Department of Molecular Microbiology, MBW/BCA, Faculty of Biology, Vrije Universiteit Amsterdam, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands
Chongju University College of Science and Engineering, Chongju City, Korea
nstitut de Génétique et Microbiologie, Université Paris Sud, URA CNRS 2225, Université Paris XI–Bâtiment 409, 91405 Orsay Cedex, France
Centro Nacional de Biotecnologia (CSIC), Campus Universidad Autonoma, Cantoblanco, 28049 Madrid, Spain
National Institute of Basic Biology, 38 Nishigounaka, Myoudaiji-chou, Okazaki 444, Japan
Gesellschaft für Analyse-Technik und Consulting mbH, Fritz-Arnold Straße 23, D-78467 Konstanz, Germany
Department of Microbiology, Faculty of Agronomy, 6 Avenue du Maréchal Juin, B-5030 Gembloux, Belgium
Biotech Research, BMF, Wilhelmsfeld, Klingelstrasse 35, D-69434 Hirschhorn, Germany
Department of Applied Biology, Faculty of Textile Science and Technology, Shinshu University 3-15-1, Tokida, Ueda-shi, Nagano 386, Japan
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan
Department of Marine Science, School of Marine Science and Technology, Tokai University, 3-20-1 Orido Shimizu, Shizuoka 424, Japan
European Commission, DG XII-E-1, SDME 8/78, Rue de la Loi 200, B-1049 Brussels, Belgium
AGOWA GmbH, Glienicker Weg 185, 12489 Berlin, Germany

acillus subtilis is the best-characterized member of the Gram-positive bacteria. Its genome of 4,214,810 base pairs mprises 4,100 protein-coding genes. Of these protein-coding genes, 53% are represented once, while a quarter of e genome corresponds to several gene families that have been greatly expanded by gene duplication, the largest mily containing 77 putative ATP-binding transport proteins. In addition, a large proportion of the genetic capacity is evoted to the utilization of a variety of carbon sources, including many plant-derived molecules. The identification of e signal peptidase genes, as well as several genes for components of the secretion apparatus, is important given the apacity of *Bacillus* strains to secrete large amounts of industrially important enzymes. Many of the genes are involved the synthesis of secondary metabolites, including antibiotics, that are more typically associated with *Streptomyces* ecies. The genome contains at least ten prophages or remnants of prophages, indicating that bacteriophage fection has played an important evolutionary role in horizontal gene transfer, in particular in the propagation of cterial pathogenesis.

Techniques for large-scale DNA sequencing have brought about a revolution in our perception of genomes. Together with our understanding of intermediary metabolism, it is now realistic to envisage a time when it should be possible to provide an extensive chemical definition of many living organisms. During the past couple of years, the genome sequences of *Haemophilus influenzae*, *Mycoplasma genitalium*, *Synechocystis* PCC6803, *Methanococcus jannaschii*, *M. pneumoniae*, *Escherichia coli*, *Helicobacter pylori*, *Archaeoglobus fulgidus* and the yeast *Saccharomyces cerevisiae* have been published in their entirety[1–8], and at least 40 prokaryotic genomes are currently being sequenced. Regularly updated lists of genome sequencing projects are available at http://www.mcs.anl.gov/home/gaasterl/genomes.html (Argonne National Laboratory, Illinois, USA) and http://www.tigr.org (TIGR, Rockville, Maryland, USA).

The list of sequenced microorganisms does not currently include a paradigm for Gram-positive bacteria, which are known to be important for the environment, medicine and industry. *Bacillus subtilis* has been chosen to fill this gap[9,10] as its biochemistry, physiology and genetics have been studied intensely for more than 40 years. *B. subtilis* is an aerobic, endospore-forming, rod-shaped bacterium commonly found in soil, water sources and in association with plants. *B. subtilis* and its close relatives are an important source of industrial enzymes (such as amylases and proteases), and much of the commercial interest in these bacteria arises from their capacity to secrete these enzymes at gram per litre concentrations. It has therefore been used for the study of protein secretion and for development as a host for the production of heterologous proteins[11]. *B. subtilis* (*natto*) is also used in the production of Natto, a traditional Japanese dish of fermented soya beans.

Under conditions of nutritional starvation, *B. subtilis* stops growing and initiates responses to restore growth by increasing metabolic diversity. These responses include the induction of motility and chemotaxis, and the production of macromolecular hydrolases (proteases and carbohydrases) and antibiotics. If these responses fail to re-establish growth, the cells are induced to form chemically, irradiation- and desiccation-resistant endospores. Sporulation involves a perturbation of the normal cell cycle and the differentiation of a binucleate cell into two cell types. The division of the cell into a smaller forespore and a larger mother cell, each with an entire copy of the chromosome, is the first morphological indication of sporulation. The former is engulfed by the latter and differential expression of their respective genomes, coupled to a complex network of interconnected regulatory pathways and developmental checkpoints, culminates in the programmed death and lysis of the mother cell and release of the mature spore[12]. In an alternative developmental process, *B. subtilis* is also able to differentiate into a physiological state, the competent state, that allows it to undergo genetic transformation[13].

General features of the DNA sequence

Analysis at the replicon level. The *B. subtilis* chromosome has 4,214,810 base pairs (bp), with the origin of replication coinciding with the base numbering start point[14], and the terminus at about 2,017 kilobases (kb)[15]. The average G + C ratio is 43.5%, but varies considerably throughout the chromosome. This average is also different if one considers the nucleotide content of coding sequences, for which G and A (24% and 30%) are relatively more abundant than their counterparts C and T (20% and 26%). A significant inversion of the relative G − C/G + C ratio is visible at the origin of replication, indicating asymmetry of the nucleotide composition between the replication leading strand and the lagging strand[16]. Several A + T-rich islands are likely to reveal the signature of bacteriophage lysogens or other inserted elements (Fig. 1; see below).

We have analysed the abundance of oligonucleotides ('words') in the genome in various ways: absolute number of words in the genomic text, or comparison with the expected count derived from several models of the chromosome (for example, Markov models, or simulated sequences in which previously known features of the genome were conserved[17]). Comparing the experimental data with various models allowed us to define under- and overrepresentation of words in the experimental data set by reference to the model chosen. In general, the dinucleotide bias follows closely what has been described for other prokaryotes[18,19], in that the dinucleotides most overrepresented are AA, TT and GC, whereas those less represented are TA, AC and GT. Plots of the frequencies of AG, GA, CT and TC in sliding windows along the chromosome show dramatic decreases or increases around the origin and terminus of replication (data not shown). Trinucleotide frequency, directly related to the coding frame, will be discussed below. The distribution of words of four, five and six nucleotides shows significant correlations between the usage of some words and replication (several such oligonucleotides are very significantly overrepresented in one of the strands and underrepresented in the other one).

Setting a statistical cut-off for the significance of duplications at 10^{-3}, we expected duplication by chance of words longer than 24 nucleotides to be rare[20]. In fact, the genome of *B. subtilis* contains a plethora of such duplications, some of them appearing more than

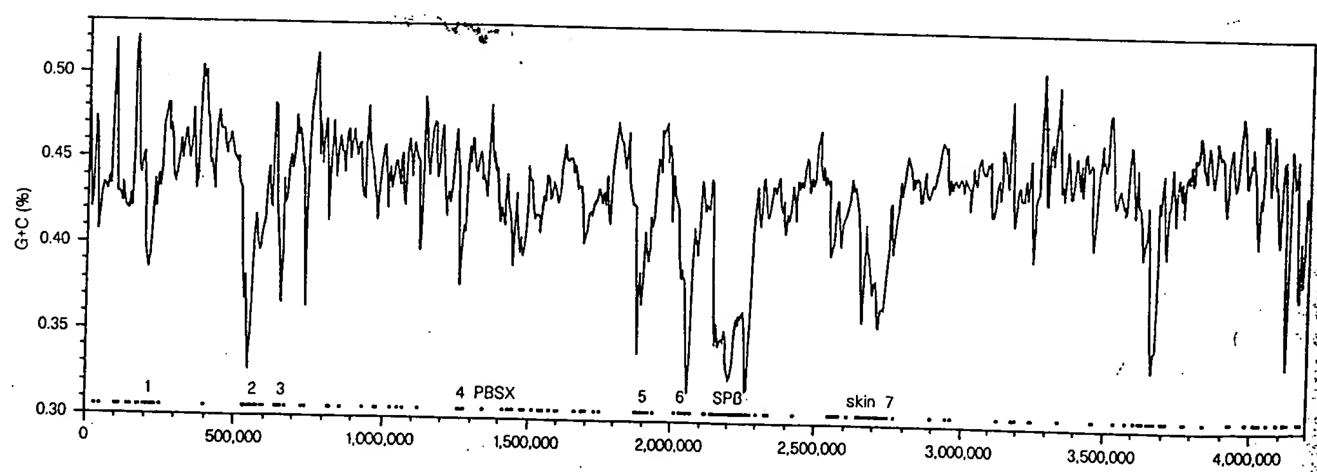


Figure 1 Distribution of A + T-rich islands along the chromosome of *B. subtilis*, in sliding windows of 10,000 nucleotides, with a step of 5,000 nucleotides. Location of genes from class 3 according to codon usage analysis (see Fig. 4) is indicated by dots at the bottom of the graph. Known prophages (PBSX, SPβ and *skin*) are indicated by their names, and prophage-like elements are numbered from 1 to 7.

bglH 4033 β-glucosidase (cellulose degradation)
bglS 4011 endo-1,3-1,4 glucanase (lichenan degradation)
crh 3569 catabolite repression HPr-like protein
csn 2748 chitosanase
csrA 3635 carbon storage regulator
fruB 1508 fructose 1-phosphate kinase
galE 3990 UDP-glucose 4-epimerase (galactose metabolism)
galK 3921 galactokinase (galactose metabolism)
galT 3919 galactose-1-phosphate uridyltransferase (galactose metabolism)
gdh 445 glucose 1-dehydrogenase
glcK 2571 glucose kinase
glgA 3167 starch (bacterial glycogen) synthase (glycogen biosynthesis)
glgB 3171 1,4-glucan branching enzyme (glycogen biosynthesis)
glgC 3169 glucose-1-phosphate adenylyltransferase (glycogen biosynthesis)
glgD 3168 required for glycogen biosynthesis
glgP 3165 glycogen phosphorylase (glycogen metabolism)
glpD 1004 glycerol-3-phosphate dehydrogenase (glycerol utilization)
glpK 1003 glycerol kinase (glycerol utilization)
glvA 890 6-phospho-α-glucosidase (arbutin fermentation)
gntK 4113 gluconate kinase (gluconate utilization)
gntZ 4116 6-phosphogluconate dehydrogenase (gluconate utilization)
gpsA 2389 NAD(P)H-dependent glycerol-3-phosphate dehydrogenase
gutB 667 sorbitol dehydrogenase
iolB 4082 myo-inositol catabolism
iolC 4081 myo-inositol catabolism
iolD 4080 myo-inositol catabolism
iolE 4078 myo-inositol catabolism
iolG 4076 myo-inositol 2-dehydrogenase (inositol catabolism)
iolH 4075 myo-inositol catabolism
iolI 4074 myo-inositol catabolism
iolS 4084 myo-inositol catabolism
kdgA 2323 deoxyphosphogluconate aldolase (pectin utilization)
kdgK 2324 2-keto-3-deoxygluconate kinase (pectin utilization)
kduD 2326 2-keto-3-deoxygluconate oxidoreductase (pectin utilization)
kduI 2325 5-keto-4-deoxyuronate isomerase (pectin utilization)
lacA 3504 β-galactosidase
lctE 329 L-lactate dehydrogenase
licH 3959 6-phospho-β-glucosidase
lplD 782 hydrolytic enzyme
melA 3100 α-D-galactoside galactohydrolase
mtlD 451 mannitol-1-phosphate dehydrogenase
nagA 3594 N-acetylglucosamine-6-phosphate deacetylase (N-acetyl glucosamine utilization)
nagB 3596 N-acetylglucosamine-6-phosphate isomerase (N-acetyl glucosamine utilization)
narQ 3773 required for formate dehydrogenase activity
pel 828 pectate lyase
pelB 2034 pectate lyase
pmi 3688 mannose-6-phosphate isomerase
pps 2053 phosphoenolpyruvate synthase
pta 3865 phosphotransacetylase
ptsH 1459 histidine-containing phosphocarrier protein of the phosphotransferase system (PTS) (HPr protein)
rbsK 3701 ribokinase (ribose metabolism)
sacA 3902 sucrase-6-phosphate hydrolase
sacB 3535 levansucrase
sacC 2759 levanase
sacX 3941 negative regulatory protein of SacY
treA 851 trehalose-6-phosphate hydrolase
xsa 2914 β-xylosidase / α-L-arabinosidase (xylan degradation)
xylA 1891 xylose isomerase (xylose metabolism)
xylB 1893 xylulose kinase (xylose metabolism)
xynA 2054 endo-1,4-xylanase (xylan degradation)
xynB 1888 xylan β-1,4-xylosidase (xylan degradation)
xynD 1945 endo-1,4-xylanase (xylan degradation)
ybaN 161 polysaccharide deacetylase
ybbD 188 β-hexosaminidase
ybcM 213 glucosamine-fructose-6-phosphate aminotransferase
ybfT 258 glucosamine-6-phosphate isomerase
ycbC 268 5-dehydro-4-deoxyglucarate dehydratase
ycbD 269 aldehyde dehydrogenase
ycbF 272 glucarate dehydratase
ycdF 305 glucose 1-dehydrogenase
ycdG 306 oligo-1,6-glucosidase
ycgS 352 aromatic hydrocarbon catabolism
yckE 370 β-glucosidase
yckG 375 D-arabino 3-hexulose 6-phosphate formaldehyde lyase
ycsN 466 aryl-alcohol dehydrogenase
ydaD 471 alcohol dehydrogenase
ydaF 473 acetyltransferase
ydaM 482 cellulose synthase
ydaP 488 pyruvate oxidase
ydhP 628 β-glucosidase
ydhR 631 fructokinase
ydhS 632 mannose-6-phosphate isomerase
ydhT 632 mannan endo-1,4-mannosidase
ydjE 670 fructokinase
ydjL 679 L-iditol 2-dehydrogenase
ydjP 682 arylesterase
yeaC 688 methanol dehydrogenase regulation
yesY 774 rhamnogalacturonan acetylesterase
yesZ 774 β-galactosidase
yfhM 929 epoxide hydrolase
yfhR 937 glucose 1-dehydrogenase
yfjS 869 polysaccharide deacetylase
yfmT 807 benzaldehyde dehydrogenase
yfnH 798 glucose-1-phosphate cytidylyltransferase
ygaK 958 reticuline oxidase
yhcW 997 phosphoglycolate phosphatase
yhdF 1022 glucose 1-dehydrogenase
yhdN 1030 aldo/keto reductase
yheN 1041 endo-1,4-xylanase
yhlE 1095 glucanase
yhxB 1006 phosphomannomutase
yhxC 1115 alcohol dehydrogenase
yhxD 1118 ribitol dehydrogenase
yisS 1164 myo-inositol 2-dehydrogenase
yitF 1175 mandelate racemase
yitY 1192 L-gulonolactone oxydase
yjdE 1274 mannose-6-phosphate isomerase
yjeA 1281 endo-1,4-xylanase
yjgC 1285 formate dehydrogenase
yjmA 1300 glucuronate isomerase
yjmD 1304 sorbitol dehydrogenase
yjmE 1305 D-mannonate hydrolase
yjmF 1306 2-deoxy-D-gluconate 3-dehydrogenase
yjmI 1309 tagaturonate reductase
yjmJ 1311 altronate hydrolase
ykcC 1356 dolichol phosphate mannose synthase
ykfB 1366 chloromuconate cycloisomerase
ykfC 1367 polysugar degrading enzyme
ykoT 1403 dolichol phosphate mannose synthase
ykrW 1427 ribulose-bisphosphate carboxylase
yktC 1537 myo-inositol-1(or 4)-monophosphatase
ykuF 1477 glucose 1-dehydrogenase
ykvO 1442 glucose 1-dehydrogenase
ykvQ 1445 chitinase
yloR 1653 ribulose-5-phosphate 3-epimerase
ylxY 1741 deacetylase
ynfF 1943 endo-xylanase
yngE 1951 propionyl-CoA carboxylase
yoaC 2023 xylulokinase
yoaD 2024 phosphoglycerate dehydrogenase
yoaE 2025 formate dehydrogenase
yoaI 2031 4-hydroxyphenylacetate-3-hydroxylase
yogA 2007 alcohol dehydrogenase
yqiQ 2507 phosphoenolpyruvate mutase
yqjD 2488 propionyl-CoA carboxylase
yrhE 2760 formate dehydrogenase
yrhG 2780 formate dehydrogenase
yrhH 2778 methyltransferase
yrhO 2768 cyclodextrin metabolism
yrpG 2742 sugar-phosphate dehydrogenase
ysdC 2950 endo-1,4-glucanase
ysfC 2932 glycolate oxidase subunit
ysfD 2934 glycolate oxidase subunit
ytbE 2969 plant metabolite dehydrogenase
ytcA 3155 NDP-sugar dehydrogenase
ytcB 3156 NDP-sugar epimerase
ytcI 3024 acetate-CoA ligase
ytdA 3155 UTP-glucose-1-phosphate uridylyltransferase
ytiB 3138 carbonic anhydrase
ytoP 3055 endo-1,4-glucanase
yttI 2989 acetyl-CoA carboxylase
yugF 3227 dihydrolipoamide S-acetyltransferase
yugJ 3224 NADH-dependent butanol dehydrogenase
yugK 3222 NADH-dependent butanol dehydrogenase
yugT 3215 exo-1,4-glucosidase
yulC 3200 rhamnulokinase
yulE 3198 L-rhamnose isomerase
yusZ 3382 retinol dehydrogenase
yutF 3318 N-acetyl-glucosamine catabolism
yuxG 3203 sorbitol-6-phosphate 2-dehydrogenase
yvaM 3455 hydrolase
yvcN 3568 N-hydroxyarylamine O-acetyltransferase
yvcT 3562 glycerate dehydrogenase
yvdA 3561 carbonic anhydrase
yvdF 3557 glucan 1,4-maltohydrolase
yvdL 3548 oligo-1,6-glucosidase
yvdM 3547 β-phosphoglucomutase
yveB 3537 levanase
yvfO 3502 arabinogalactan endo-1,4-galactosidase
yvfQ 3499 hydrolase
yvfV 3495 glycolate oxidase
yvgN 3427 plant-metabolite dehydrogenase
yvkC 3615 pyruvate,water dikinase
yvoE 3592 phosphoglycolate phosphatase
yvoF 3591 O-acetyltransferase
yvpA 3590 pectate lyase
yvyH 3664 UDP-N-acetylglucosamine 2-epimerase
ywdH 3895 aldehyde dehydrogenase
ywfD 3872 glucose 1-dehydrogenase
ywjI 3805 glycerol-inducible protein
ywqF 3730 NDP-sugar dehydrogenase
yxbG 4091 glucose 1-dehydrogenase
yxiA 4040 arabinan endo-1,5-L-arabinosidase
yxjF 4000 gluconate 5-dehydrogenase
yxnA 4107 glucose 1-dehydrogenase
yyaE 4202 formate dehydrogenase
yyaI 4196 galactoside acetyltransferase
yycR 4136 formaldehyde dehydrogenase

II.1.2 MAIN GLYCOLYTIC PATHWAYS 28

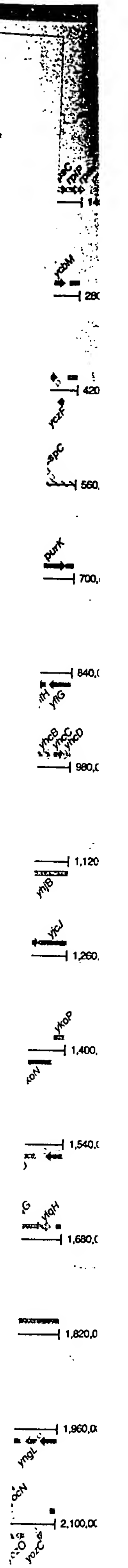
eno 3477 enolase (glycolysis)
fbaA 3808 fructose-1,6-bisphosphate aldolase (glycolysis)
fbp 4127 fructose-1,6-bisphosphatase (gluconeogenesis)
gap 3482 glyceraldehyde 3-phosphate dehydrogenase (glycolysis)
gapB 2967 glyceraldehyde 3-phosphate dehydrogenase (glycolysis)
iolJ 4073 fructose-1,6-bisphosphate aldolase (glycolysis)
pckA 3129 phosphoenolpyruvate carboxykinase
pdhA 1528 pyruvate dehydrogenase (E1 α subunit)
pdhB 1529 pyruvate dehydrogenase (E1 β subunit)
pdhC 1530 pyruvate dehydrogenase (dihydrolipoamide acetyltransferase E2 subunit)
pdhD 1531 pyruvate dehydrogenase / 2-oxoglutarate dehydrogenase (dihydrolipoamide dehydrogenase E3 subunit)
pfk 2987 6-phosphofructokinase (glycolysis)
pgi 3221 glucose-6-phosphate isomerase (glycolysis)
pgk 3480 phosphoglycerate kinase (glycolysis)
pgm 3478 phosphoglycerate mutase (glycolysis)
pycA 1554 pyruvate carboxylase
pykA 2986 pyruvate kinase (glycolysis)
tkt 1919 transketolase (pentose phosphate)
tpi 3479 triose phosphate isomerase (glycolysis)
ybbT 198 phosphoglucomutase (glycolysis)
ydeA 558 glyceraldehyde 3-phosphate dehydrogenase (glycolysis)
yhfR 1109 phosphoglycerate mutase (glycolysis)
yqeC 2651 6-phosphogluconate dehydrogenase (pentose phosphate)
yqiV 2501 dihydrolipoamide dehydrogenase
yqjI 2481 6-phosphogluconate dehydrogenase (pentose phosphate)
yqjJ 2478 glucose-6-phosphate 1-dehydrogenase (pentose phosphate)
ywjH 3807 transaldolase (pentose phosphate)
ywlF 3791 ribose 5-phosphate epimerase (pentose phosphate)

II.1.3 TCA CYCLE 19

citA 1021 citrate synthase I
citB 1926 aconitate hydratase
citC 2980 isocitrate dehydrogenase
citG 3389 fumarate hydratase
citH 2979 malate dehydrogenase
citZ 2981 citrate synthase II
maiS 3058 malate dehydrogenase
mmgD 2510 citrate synthase III
odhA 2111 2-oxoglutarate dehydrogenase (E1 subunit)
odhB 2108 2-oxoglutarate dehydrogenase (dihydrolipoamide transsuccinylase, E2 subunit)
sdhA 2907 succinate dehydrogenase (flavoprotein subunit)
sdhB 2905 succinate dehydrogenase (iron-sulphur protein)
sdhC 2908 succinate dehydrogenase (cytochrome b_{558} subunit)
sucC 1680 succinyl-CoA synthetase (β subunit)
sucD 1681 succinyl-CoA synthetase (α subunit)
yjmC 1303 malate dehydrogenase
yqkJ 2452 malate dehydrogenase
ytsJ 2990 malate dehydrogenase
ywkA 3801 malate dehydrogenase

II.2 METABOLISM OF AMINO ACIDS AND RELATED MOLECULES 205

ald 3277 L-alanine dehydrogenase
ampS 1516 aminopeptidase
ansA 2456 L-asparaginase
ansB 2455 L-aspartase
aprE 1105 extracellular alkaline serine protease (subtilisin E)
aprX 1862 intracellular alkaline serine protease
argB 1197 N-acetylglutamate 5-phosphotransferase (arginine biosynthesis)
argC 1195 N-acetylglutamate γ-semialdehyde dehydrogenase (arginine biosynthesis)
argD 1198 N-acetylornithine aminotransferase (arginine biosynthesis)
argE 2142 acetylornithine deacetylase (arginine biosynthesis)
argF 1203 ornithine carbamoyltransferase (arginine biosynthesis)
argG 3013 argininosuccinate synthase (arginine biosynthesis)
argH 3012 argininosuccinate lyase (arginine biosynthesis)
argJ 1196 ornithine acetyltransferase / amino-acid acetyltransferase (arginine biosynthesis)
aroA 3046 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase / chorismate mutase-isozyme 3 (shikimate pathway)
aroB 2378 3-dehydroquinate synthase (shikimate pathway)
aroC 2413 3-dehydroquinate dehydratase (shikimate pathway)
aroD 2645 shikimate 5-dehydrogenase (shikimate pathway)
aroE 2368 5-enolpyruvoylshikimate-3-phosphate synthase (shikimate pathway)
aroF 2380 chorismate synthase (shikimate pathway)
aroH 2377 chorismate mutase (isozymes 1 and 2) (aromatic amino acids biosynthesis)
aroI 340 shikimate kinase (shikimate pathway)
asd 1745 aspartate-semialdehyde dehydrogenase
ask 2910 aspartokinase II attenuator
asnB 3127 asparagine synthetase
asnH 4098 asparagine synthetase
aspB 2348 aspartate aminotransferase
bcsA 2317 naringenin-chalcone synthase (phenylalanine metabolism)
bfmBAA 2499 branched-chain α-keto acid dehydrogenase E1 (2-oxoisovalerate dehydrogenase α subunit)
bfmBAB 2498 branched-chain α-keto acid dehydrogenase E1 (2-oxoisovalerate dehydrogenase β subunit)
bfmBB 2497 branched-chain α-keto acid dehydrogenase E2 subunit (lipoamide acyltransferase)
bltD 2718 spermine/spermidine acetyltransferase
bpr 1599 bacillopeptidase F
cad 1535 lysine decarboxylase
carA 1199 carbamoyl-phosphate transferase-arginine (subunit A) (arginine biosynthesis)
carB 1200 carbamoyl-phosphate transferase-arginine (subunit B) (arginine biosynthesis)
ctpA 2133 carboxy-terminal processing protease
cysE 113 serine acetyltransferase (cysteine biosynthesis)
cysH 1630 phosphoadenosine phosphosulfate reductase (cysteine biosynthesis)
cysK 82 cysteine synthetase A (cysteine biosynthesis)
dal 517 D-alanine racemase
dapA 1748 dihydrodipicolinate synthase (diaminopimelate/lysine biosynthesis)
dapB 2359 dihydrodipicolinate reductase (diaminopimelate/lysine biosynthesis)
dapG 1747 aspartokinase I (α and β subunits)
def 1646 polypeptide deformylase
epr 3939 minor extracellular serine protease
glmS 200 L-glutamine-D-fructose-6-phosphate amidotransferase
glnA 1878 glutamine synthetase
gltA 2014 glutamate synthase (large subunit) (glutamate biosynthesis)
gltB 2009 glutamate synthase (small subunit) (glutamate biosynthesis)
glyA 3789 serine hydroxymethyltransferase (glycine/serine/threonine metabolism)
hisA 3584 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (histidine biosynthesis)
hisB 3585 imidazoleglycerol-phosphate dehydratase (histidine biosynthesis)
hisC 2371 histidinol-phosphate aminotransferase (histidine biosynthesis) / tyrosine and phenylalanine aminotransferase
hisD 3587 histidinol dehydrogenase (histidine biosynthesis)
hisF 3583 HisF cyclase-like protein (synthesis of D-erythro-imidazole glycerol phosphate)
hisG 3587 ATP phosphoribosyltransferase (histidine biosynthesis)
hisH 3585 amidotransferase (histidine biosynthesis)
hisI 3583 phosphoribosyl-AMP cyclohydrolase / phosphoribosyl-ATP pyrophosphohydrolase (histidine biosynthesis)
hom 3315 homoserine dehydrogenase (threonine/methionine biosynthesis)
hutG 4045 formiminoglutamate hydrolase (histidine utilization)
hutH 4041 histidase (histidine utilization)
hutI 4044 imidazolone-5-propionate hydrolase (histidine utilization)
hutU 4042 urocanase (histidine utilization)
ilvA 2293 threonine dehydratase (isoleucine biosynthesis)
ilvB 2896 acetolactate synthase (large subunit) (valine/isoleucine biosynthesis)
ilvC 2894 ketol-acid reductoisomerase (valine/isoleucine biosynthesis)
ilvD 2302 dihydroxy-acid dehydratase (valine/isoleucine biosynthesis)
ilvN 2894 acetolactate synthase (small subunit)



flgE 1700 flagellar hook protein
flgK 3639 flagellar hook-associated protein 1 (HAP1)
flgL 3637 flagellar hook-associated protein 3 (HAP3)
flgM 3640 flagellin synthesis regulatory protein (anti-sigma factor [σ^D])
flhA 1707 flagella-associated protein
flhB 1706 flagella-associated protein
flhF 1709 flagella-associated protein
flhO 3746 flagellar basal-body rod protein
flhP 3745 flagellar hook-basal body protein
fliD 3633 flagellar hook-associated protein 2 (HAP2)
fliE 1692 flagellar hook-basal body protein
fliF 1692 flagellar basal-body M-ring protein
fliG 1694 flagellar motor switch protein
fliH 1695 flagellar assembly protein
fliI 1695 flagellar-specific ATP synthase
fliJ 1697 flagellar protein required for formation of basal body
fliK 1698 flagellar hook-length control
fliL 1701 flagellar protein required for flagellar formation
fliM 1701 flagellar motor switch protein
fliP 1704 flagellar protein required for flagellar formation
fliQ 1705 flagellar protein required for flagellar formation
fliR 1705 flagellar protein required for flagellar formation
fliS 3632 flagellar protein
fliT 3632 flagellar protein
fliY 1702 flagellar motor switch protein
fliZ 1704 flagellar protein required for flagellar formation
hag 3635 flagellin protein
mcpA 3207 methyl-accepting chemotaxis protein (glucose and α-methyl-glucoside)
mcpB 3212 methyl-accepting chemotaxis protein (asparagine, glutamine and histidine)
mcpC 1463 methyl-accepting chemotaxis protein (cysteine, proline, threonine, glycine, serine, lysine, valine and arginine)
motA 1435 motility protein (flagellar motor rotation)
motB 1434 motility protein (flagellar motor rotation)
tlpA 3209 methyl-accepting chemotaxis protein
tlpB 3205 methyl-accepting chemotaxis protein
tlpC 374 methyl-accepting chemotaxis protein
yfmS 808 methyl-accepting chemotaxis protein
yhfV 1113 methyl-accepting chemotaxis protein
ylqH 1679 flagellar biosynthetic protein
ylxG 1699 flagellar hook assembly protein
ylxH 1710 flagellar biosynthesis switch protein
yoaH 2030 methyl-accepting chemotaxis protein
ytxD 3043 flagellar motor apparatus
ytxE 3042 motility protein
yvaQ 3457 transmembrane receptor taxis protein
yvyC 3634 flagellar protein
yvyF 3640 flagellar protein
yvyG 3639 flagellar protein
yvzB 3609 flagellin

csaA 2079 chaperonin involved in protein secretion
ffh 1672 signal recognition particle
ftsY 1670 signal recognition particle
lsp 1616 signal peptidase II
lytA 3662 secretion of major autolysin LytC
prsA 1071 protein secretion (post-translocation chaperonin)
secA 3630 preprotein translocase subunit
secE 118 preprotein translocase subunit
secF 2828 protein-export membrane protein (product also similar to SecD of *E. coli*)
secY 145 preprotein translocase subunit
sipS 2432 signal peptidase I
sipT 1511 signal peptidase I
sipU 454 signal peptidase I
sipV 1122 signal peptidase I
sipW 2554 signal peptidase I
yaaT 42 signal peptidase II
yacD 81 protein secretion PrsA homologue
yobE 2057 general secretion pathway protein

divIB 1593 cell-division initiation protein (septum formation)
divIC 69 cell-division initiation protein (septum formation)
divIVA 1612 cell-division initiation protein (septum placement)
ftsA 1596 cell-division protein (septum formation)
ftsE 3625 cell-division ATP-binding protein
ftsH 77 cell-division protein / general stress protein (class III heat-shock)
ftsL 1581 cell-division protein (septum formation)
ftsX 3624 cell-division protein
ftsZ 1597 cell-division initiation protein (septum formation)
gid 1685 glucose-inhibited division protein
gidA 4211 glucose-inhibited division protein
gidB 4209 glucose-inhibited division protein
maf 2862 septum formation
minC 2859 cell-division inhibitor (septum placement)
minD 2858 cell-division inhibitor (septum placement) (ATPase activator of MinC)
yacA 75 cell-cycle protein
ylhF 925 cell-division inhibitor
yjoB 1314 cell-division protein FtsH homologue
ylaO 1552 cell-division protein
ylmH 1611 cell-division protein
ywcF 3912 cell-division protein

bofA 30 inhibitor of the pro-σ^K processing machinery
bofC 2837 forespore regulator of the σ^K checkpoint
cgeA 2148 maturation of the outermost layer of the spore
cgeB 2148 maturation of the outermost layer of the spore
cgeC 2148 maturation of the outermost layer of the spore
cgeD 2147 maturation of the outermost layer of the spore
cgeE 2146 maturation of the outermost layer of the spore
cotA 685 spore coat protein (outer)
cotB 3715 spore coat protein (outer)
cotC 1905 spore coat protein (outer)
cotD 2332 spore coat protein (inner)
cotE 1774 spore coat protein (outer)
cotF 4166 spore coat protein
cotG 3716 spore coat protein
cotH 3716 spore coat protein (inner)
cotJA 755 polypeptide composition of the spore coat
cotJB 756 polypeptide composition of the spore coat
cotJC 756 polypeptide composition of the spore coat
cotK 1926 spore coat protein
cotL 1926 spore coat protein
cotM 1925 spore coat protein (outer)
cotN 2553 spore coat-associated protein
cotS 3160 spore coat protein
cotT 1280 spore coat protein (inner)
cotV 1251 spore coat protein (insoluble fraction)
cotW 1251 spore coat protein (insoluble fraction)
cotX 1251 spore coat protein (insoluble fraction)
cotY 1250 spore coat protein (insoluble fraction)
cotZ 1249 spore coat protein (insoluble fraction)
csgA 228 sporulation-specific SASP protein
jag 4213 SpoIIIJ-associated protein
kapB 3230 activator of KinB in the initiation of sporulation
kapD 3232 inhibitor of the KinA pathway to sporulation
kbaA 159 activation of the KinB signaling pathway to sporulation
obg 2853 GTP-binding protein involved in initiation of sporulation (Spo0A activation)
phrA 1316 phosphatase (RapA) inhibitor (imported by Opp)
phrC 430 phosphatase (RapC) regulator / competence and sporulation stimulating factor (CSF)
phrE 2660 phosphatase (RapE) regulator
phrF 3846 phosphatase (RapF) regulator
phrG 4141 phosphatase (RapG) regulator
phrI 543 phosphatase (RapI) regulator
phrK 2063 phosphatase (RapK) regulator
rapA 1315 response regulator aspartate phosphatase [Spo0F~P]
rapB 3771 response regulator aspartate phosphatase [Spo0F~P]
rapC 428 response regulator aspartate phosphatase
rapD 3743 response regulator aspartate phosphatase
rapE 2558 response regulator aspartate phosphatase
rapF 3845 response regulator aspartate phosphatase
rapG 4139 response regulator aspartate phosphatase
rapH 750 response regulator aspartate phosphatase
rapI 547 response regulator aspartate phosphatase
rapJ 304 response regulator aspartate phosphatase
rapK 2061 response regulator aspartate phosphatase
sinI 2552 antagonist of SinR
soj 4206 centromere-like function involved in forespore chromosome partitioning / inhibition of Spo0A activation
splB 1461 spore photoproduct lyase
spmA 2423 spore maturation protein (spore core dehydratation)
spmB 2422 spore maturation protein (spore core dehydratation)
spo0B 2854 sporulation initiation phosphotransferase (part of phosphorelay: Spo0F~P->Spo0B~P->Spo0A~P)
spo0E 1430 negative sporulation regulatory phosphatase [Spo0A~P]
spo0J 4206 chromosome positioning near the pole and transport through the polar septum / antagonist of Soj
spoIIAA 2444 anti-anti-sigma factor [SpoIIAB]
spoIIAB 2444 anti-sigma factor [σ^F(SpoIIAC)] and serine kinase [SpoIIAA]
spoIIB 2864 endospore development (oligosporogenous mutation)
spoIID 3777 required for complete dissolution of the asymmetric septum
spoIIE 71 serine phosphatase [SpoIIAA~P] (σ^F activation) / asymmetric septum formation
spoIIGA 1603 protease (processing of pro-σ^E to active σ^E)
spoIIIAA 2537 mutants block sporulation after engulfment
spoIIIAB 2536 mutants block sporulation after engulfment
spoIIIAC 2535 mutants block sporulation after engulfment
spoIIIAD 2535 mutants block sporulation after engulfment
spoIIIAE 2535 mutants block sporulation after engulfment
spoIIIAF 2534 mutants block sporulation after engulfment
spoIIIAG 2533 mutants block sporulation after engulfment
spoIIIAH 2532 mutants block sporulation after engulfment
spoIIIE 1752 DNA translocase required for chromosome partitioning through the septum into the forespore
spoIIIJ 4214 essential for σ^G activity at stage III
spoIIM 2450 required for dissolution of the septal cell wall
spoIIP 2534 required for dissolution of the septal cell wall
spoIIQ 3760 required for completion of engulfment
spoIIR 3794 required for processing of pro-σ^E
spoIISA 1349 lethal when synthesized during vegetative growth in the absence of SpoIISB
spoIISB 1348 disruption blocks sporulation after septum formation
spoIVA 2387 required for proper spore cortex formation and coat assembly
spoIVB 2520 intercompartmental signalling of pro-σ^K processing/activation in the mother-cell
spoIVCA 2654 site-specific DNA recombinase required for creating the *sigK* gene (excision of the *skin* element)
spoIVFA 2857 inhibitor of SpoIVFB
spoIVFB 2856 protease (processing of pro-σ^K to active σ^K)
spoVAA 2443 mutants lead to the production of immature spores
spoVAB 2442 mutants lead to the production of immature spores
spoVAC 2441 mutants lead to the production of immature spores
spoVAD 2441 mutants lead to the production of immature spores
spoVAE 2440 mutants lead to the production of immature spores
spoVAF 2439 mutants lead to the production of immature spores
spoVB 2829 involved in spore cortex synthesis
spoVC 60 thermosensitive mutant blocks spore coat formation
spoVE 1590 required for spore cortex synthesis
spoVFA 1744 dipicolinate synthase subunit A
spoVFB 1745 dipicolinate synthase subunit B
spoVG 56 required for spore cortex synthesis
spoVID 2872 required for assembly of the spore coat
spoVK 1873 disruption leads to the production of immature spores
spoVM 1655 required for normal spore cortex and coat synthesis
spoVR 1015 involved in spore cortex synthesis
spoVS 1769 required for dehydratation of the spore core and assembly of the coat
spsA 3892 spore coat polysaccharide synthesis
spsB 3891 spore coat polysaccharide synthesis
spsC 3890 spore coat polysaccharide synthesis
spsD 3889 spore coat polysaccharide synthesis
spsE 3888 spore coat polysaccharide synthesis
spsF 3887 spore coat polysaccharide synthesis
spsG 3886 spore coat polysaccharide synthesis
spsI 3885 spore coat polysaccharide synthesis
spsJ 3884 spore coat polysaccharide synthesis
spsK 3883 spore coat polysaccharide synthesis
sspA 3025 small acid-soluble spore protein (major α-type SASP)
sspB 1050 small acid-soluble spore protein (major β-type SASP)
sspC 2155 small acid-soluble spore protein (minor α/β-type SASP)
sspD 1413 small acid-soluble spore protein (minor α/β-type SASP)
sspE 937 small acid-soluble spore protein (major γ-type SASP)
sspF 53 small acid-soluble spore protein (minor α/β-type SASP)
usd 3748 required for translation of *spoIIID*
yknT 1495 sporulation protein σ^E-controlled
ykvU 1449 spore cortex membrane protein
ynzH 1901 spore coat protein
yobW 2083 membrane protein σ^E-controlled
yqgT 2568 γ-D-glutamyl-L-diamino acid endopeptidase I
yqjG 2483 lipoprotein SpoIIIJ-like
yraD 2754 spore coat protein
yraE 2754 spore coat protein
yraF 2752 spore coat protein
yraG 2752 spore coat protein
yrbA 2845 spore coat protein
yrbB 2844 spore coat protein
yrbC 2843 spore coat protein
ytaA 3161 spore coat protein
ytgP 3074 spore cortex protein
ytpT 3051 DNA translocase stage III sporulation protein
yyaA 4208 DNA-binding protein Spo0J-like

gerAA 3390 germination response to L-alanine
gerAB 3391 germination response to L-alanine
gerAC 3392 germination response to L-alanine
gerBA 3688 germination response to the combination of glucose, fructose, L-asparagine, and KCl
gerBB 3689 germination response to the combination of glucose, fructose, L-asparagine, and KCl
gerBC 3690 germination response to the combination of glucose, fructose, L-asparagine, and KCl
gerCA 2384 heptaprenyl diphosphate synthase component I (menaquinone biosynthesis)
gerCB 2383 menaquinone biosynthesis methyltransferase (menaquinone biosynthesis)
gerCC 2382 heptaprenyl diphosphate synthase component II (menaquinone biosynthesis)
gerD 159 germination response to L-alanine and to the combination of glucose, fructose, L-asparagine, and KCl
gerKA 420 germination response to the combination of glucose, fructose, L-asparagine, and KCl
gerKB 423 germination response to the combination of glucose, fructose, L-asparagine, and KCl
gerKC 421 germination response to the combination of glucose, fructose, L-asparagine, and KCl
gerM 2902 germination (cortex hydrolysis) and sporulation (stage II, multiple polar septa)
gpr 2635 spore protease (degradation of SASPs)
sleB 2399 spore cortex-lytic enzyme
yfkQ 850 spore germination response
yfkR 848 spore germination protein
yfkT 847 spore germination protein
ykvT 1448 spore cortex-lytic enzyme
yndD 1907 spore germination protein
yndE 1908 spore germination protein
yndF 1909 spore germination protein

cinA 1763 competence-damage inducible protein
comC 2864 late competence protein required for processing and translocation of ComGC
comEA 2640 late competence operon required for DNA binding and uptake
comEB 2640 late competence operon required for DNA binding and uptake
comEC 2639 late competence operon required for DNA binding and uptake
comER 2640 non-essential gene for competence
comFA 3643 late competence protein required for DNA uptake
comFB 3641 late competence gene
comFC 3641 late competence gene
comGA 2559 late competence gene
comGB 2558 DNA transport machinery
comGC 2557 exogenous DNA-binding
comGD 2557 DNA transport machinery
comGE 2557 DNA transport machinery
comGF 2556 DNA transport machinery
comGG 2556 DNA transport machinery
comS 390 assembly link between regulatory components of the competence signal transduction pathway
comX 3255 competence pheromone precursor (activation of ComA)
mecA 1229 negative regulator of competence
ypbH 2403 negative regulation of competence MecA homologue

abfA 2939 α-L-arabinofuranosidase
abnA 2949 arabinan-endo 1,5-α-L-arabinase (degradation of plant cell wall polysaccharide)
ackA 3015 acetate kinase
acoA 879 acetoin dehydrogenase E1 component (TPP-dependent α subunit)
acoB 880 acetoin dehydrogenase E1 component (TPP-dependent β subunit)
acoC 881 acetoin dehydrogenase E2 component (dihydrolipoamide acetyltransferase)
acoL 882 acetoin dehydrogenase E3 component (dihydrolipoamide dehydrogenase)
acsA 3039 acetyl-CoA synthetase
acuA 3039 acetoin utilization
acuB 3040 acetoin utilization
acuC 3040 acetoin utilization
adhA 2756 NADP-dependent alcohol dehydrogenase
adhB 2753 alcohol dehydrogenase
aldX 4093 aldehyde dehydrogenase
aldY 3985 aldehyde dehydrogenase
alsD 3709 α-acetolactate decarboxylase (acetoin biosynthesis)
alsS 3710 α-acetolactate synthase (acetoin biosynthesis)
amyE 327 α-amylase
amyX 3063 pullulanase
araA 2948 L-arabinose isomerase (L-arabinose utilization)
araB 2946 L-ribulokinase (L-arabinose utilization)
araD 2945 L-ribulose-5-phosphate 4-epimerase (L-arabinose utilization)
araL 2944 L-arabinose operon
araM 2943 L-arabinose operon
bglA 4122 6-phospho-β-glucosidase
bglC 1940 endo-1,4-β-glucanase (cellulose degradation)

ydeR	578	antibiotic resistance protein
ydfA	580	arsenical pump membrane protein
ydfJ	589	antibiotic transport-associated protein
ydfL	595	multidrug-efflux transporter regulator
ydfM	596	cation efflux system
ydfO	597	ABC transporter (binding protein)
ydgF	608	amino acid ABC transporter (permease)
ydgH	609	transporter
ydgK	613	bicyclomycin resistance protein
ydhL	626	chloramphenicol resistance protein
ydhM	626	cellobiose phosphotransferase system enzyme II
ydhN	627	cellobiose phosphotransferase system enzyme II
ydhO	627	cellobiose phosphotransferase system enzyme II
ydiF	646	ABC transporter (ATP-binding protein)
ydjD	668	H^+-symporter
ydjK	676	sugar transporter
yeaB	687	cation efflux system membrane protein
yecA	712	amino acid permease
yesO	761	sugar-binding protein
yesP	762	lactose permease
yesQ	763	lactose permease
yfhA	921	iron(III) dicitrate transport permease
yfhI	926	antibiotic resistance protein
yfiB	893	ABC transporter (ATP-binding protein)
yfiC	895	ABC transporter (ATP-binding protein)
yfiG	900	metabolite transport protein
yfiL	905	ABC transporter (ATP-binding protein)
yfiM	906	ABC transporter (ATP-binding protein)
yfiN	907	ABC transporter (ATP-binding protein)
yfiS	913	multidrug resistance protein
yfiU	916	multidrug-efflux transporter
yfiY	920	iron(III) dicitrate transport permease
yfiZ	920	iron(III) dicitrate transport permease
yfjO	872	divalent cation transport protein
yfkE	865	H^+/Ca^{2+} exchanger
yfkF	865	multidrug-efflux transporter
yfkH	862	transporter
yfkL	861	multidrug resistance protein
yflA	844	aminoacid carrier protein
yflE	844	anion-binding protein
yflF	840	phosphotransferase system enzyme II
yflS	829	2-oxoglutarate/malate translocator
yfmC	826	ferrichrome ABC transporter (binding protein)
yfmD	825	ferrichrome ABC transporter (permease)
yfmE	824	ferrichrome ABC transporter (permease)
yfmF	823	ferrichrome ABC transporter (ATP-binding protein)
yfmM	815	ABC transporter (ATP-binding protein)
yfmO	812	multidrug-efflux transporter
yfmR	809	ABC transporter (ATP-binding protein)
yfnA	806	metabolite transporter
ygaD	939	ABC transporter (ATP-binding protein)
ygaL	961	nitrate ABC transporter (binding protein)
ygaM	963	ABC transporter (permease)
ygbA	962	ABC transporter (binding lipoprotein)
yhaO	1062	ABC transporter (ATP-binding protein)
yhaU	1060	Na^+/H^+ antiporter
yhcA	977	multidrug resistance protein
yhcG	981	glycine betaine/L-proline transport
yhcH	982	ABC transporter (ATP-binding protein)
yhcJ	984	ABC transporter (binding lipoprotein)
yhcL	986	sodium-glutamate symporter
yhdG	1023	amino acid transporter
yhdH	1024	sodium-dependent transporter
yheH	1047	ABC transporter (ATP-binding protein)
yheI	1045	ABC transporter (ATP-binding protein)
yheL	1044	Na^+/H^+ antiporter
yhfQ	1107	iron(III) dicitrate-binding protein
yhjB	1120	metabolite permease
yhjO	1133	multidrug-efflux transporter
yhjP	1133	transporter binding protein
yitG	1177	multidrug resistance protein
yitZ	1194	multidrug resistance protein
yjbQ	1240	Na^+/H^+ antiporter
yjdD	1272	fructose phosphotransferase system enzyme II
yjkB	1296	amino acid ABC transporter (ATP-binding protein)
yjmB	1301	Na^+-galactoside symporter
yjmG	1307	hexuronate transporter
ykaB	1350	low-affinity inorganic phosphate transporter
ykbA	1352	amino acid permease
ykcA	1353	ABC transporter (binding protein)
ykfD	1368	oligopeptide ABC transporter (permease)
yknU	1499	ABC transporter (ATP-binding protein)
yknV	1501	ABC transporter (ATP-binding protein)
yknY	1505	ABC transporter (ATP-binding protein)
ykoD	1390	cation ABC transporter (ATP-binding protein)
ykoK	1395	Mg^{2+} transporter
ykpA	1512	ABC transporter (ATP-binding protein)
ykrM	1416	Na^+-transporting ATP synthase
ykuC	1476	macrolide-efflux protein
ykvW	1451	heavy metal-transporting ATPase
ylmA	1606	ABC transporter (ATP-binding protein)
ylnA	1630	anion permease
yloB	1637	calcium-transporting ATPase
ynaJ	1387	H^+-symporter
yncC	1896	metabolite transport protein
yocN	2098	permease
yocR	2106	sodium-dependent transporter
yocS	2106	sodium-dependent transporter
yodE	2129	aromatic metabolite transporter
yodF	2130	proline permease
yojA	2125	gluconate permease
ypqE	2337	phosphotransferase system enzyme II
yqeW	2620	Na^+/P_i cotransporter
yqgG	2581	phosphate ABC transporter (binding protein)
yqgH	2580	phosphate ABC transporter (permease)
yqgI	2579	phosphate ABC transporter (permease)
yqgJ	2578	phosphate ABC transporter (ATP-binding protein)
yqgK	2577	phosphate ABC transporter (ATP-binding protein)
yqiH	2515	lipoprotein
yqiX	2492	amino acid ABC transporter (binding protein)
yqiY	2491	amino acid ABC transporter (permease)
yqiZ	2491	amino acid ABC transporter (ATP-binding protein)
yqjV	2466	multidrug resistance protein
yqkI	2453	Na^+/H^+ antiporter
yraO	2745	citrate transporter
yrbD	2841	sodium/proton-dependent alanine carrier protein
ytbD	2968	antibiotic resistance protein
ytcP	3087	ABC transporter (permease)
ytcQ	3086	lipoprotein
yteQ	3082	sugar transport protein
ytgA	3145	ABC transporter (membrane protein)
ytgB	3144	ABC transporter (ATP-binding protein)
ytgC	3143	ABC transporter (membrane protein)
ythP	3071	ABC transporter (ATP-binding protein)
ytlC	3132	anion transport ABC transporter (ATP-binding protein)
ytlD	3133	ABC transporter (permease)
ytlP	3065	ABC transporter (permease)
ytmJ	3007	amino acid ABC transporter (binding protein)
ytmK	3006	amino acid ABC transporter (binding protein)
ytmL	3006	amino acid ABC transporter (permease)
ytmM	3005	amino acid ABC transporter (permease)
ytnA	3125	proline permease
ytrB	3118	ABC transporter (ATP-binding protein)
ytrE	3115	ABC transporter (ATP-binding protein)
ytsC	3111	ABC transporter (ATP-binding protein)
ytsD	3110	ABC transporter (permease)
yttB	3108	multidrug resistance protein
yubD	3192	multidrug resistance protein
yubG	3188	Na^+-transporting ATP synthase
yufN	3239	ABC transporter (lipoprotein)
yufO	3240	ABC transporter (ATP-binding protein)
yufR	3244	organic acid transport protein
yufU	3248	Na^+/H^+ antiporter
yufV	3249	Na^+/H^+ antiporter
yugO	3218	potassium channel protein
yunJ	3330	purine permease
yunK	3331	purine permease
yurJ	3345	multiple sugar ABC transporter (ATP-binding protein)
yurM	3348	sugar permease
yurN	3349	sugar permease
yurO	3350	multiple sugar-binding protein
yurY	3360	ABC transporter (ATP-binding protein)
yusC	3363	ABC transporter (ATP-binding protein)
yusP	3374	multidrug-efflux transporter
yusV	3379	iron(III) dicitrate transport permease
yutK	3307	Na^+/nucleoside cotransporter
yuxJ	3232	multidrug-efflux transporter
yvaE	3448	multidrug-efflux transporter
yvbW	3490	amino acid permease
yvcC	3579	ABC transporter (ATP-binding protein)
yvcR	3565	ABC transporter (ATP-binding protein)
yvcS	3565	ABC transporter (permease)
yvdB	3561	transporter
yvdG	3555	maltose/maltodextrin-binding protein
yvdH	3554	maltodextrin transport system permease
yvdI	3552	maltodextrin transport system permease
yveA	3538	permease
yvfH	3510	L-lactate permease
yvfK	3508	maltose/maltodextrin-binding protein
yvfL	3506	maltodextrin transport system permease
yvfM	3505	maltodextrin transport system permease
yvfR	3498	ABC transporter (ATP-binding protein)
yvgK	3424	molybdenum-binding protein
yvgL	3424	molybdate-binding protein
yvgM	3425	molybdenum transport permease
yvgW	3440	heavy metal-transporting ATPase
yvgX	3443	heavy metal-transporting ATPase
yvgY	3443	mercuric transport protein
yvkA	3618	multidrug-efflux transporter
yvmA	3605	transporter
yvqJ	3399	macrolide-efflux protein
yvrA	3402	iron transport system
yvrB	3403	iron permease
yvrC	3403	iron-binding protein
yvrO	3413	amino acid ABC transporter (ATP-binding protein)
yvsH	3420	ABC transporter (amino acid permease)
ywbA	3938	phosphotransferase system enzyme II
ywbF	3933	sugar permease
ywcA	3923	Na^+-dependent symport
ywcJ	3904	nitrite transporter
ywfA	3874	chloramphenicol resistance
ywfF	3869	efflux protein
ywhO	3837	ABC transporter (ATP-binding protein)
ywjA	3821	ABC transporter (ATP-binding protein)
ywoA	3758	bacteriocin transport permease
ywoD	3754	transporter
ywoE	3753	permease
ywoG	3749	antibiotic resistance protein
ywpC	3743	large conductance mechanosensitive channel protein
ywrA	3721	chromate transport protein
ywrB	3720	chromate transport protein
ywrK	3712	arsenical pump membrane protein
ywtG	3693	metabolite transport protein
yxaM	4100	antibiotic resistance protein
yxcC	4087	metabolite transport protein
yxdL	4070	ABC transporter (ATP-binding protein)
yxdM	4069	ABC transporter (permease)
yxeB	4066	ABC transporter (binding protein)
yxeM	4059	amino acid ABC transporter (binding protein)
yxeN	4058	amino acid ABC transporter (permease)
yxeO	4058	amino acid ABC transporter (ATP-binding protein)
yxeR	4054	ethanolamine transporter
yxiO	4009	Mg^{2+}/citrate complex transporter
yxjA	4005	pyrimidine nucleoside transport
yxkJ	3979	metabolite-sodium symport
yxlA	3970	purine-cytosine permease
yxlF	3968	ABC transporter (ATP-binding protein)
yxlH	3966	multidrug-efflux transporter
yyaJ	4194	transporter
yybF	4180	antibiotic resistance protein
yybJ	4175	ABC transporter (ATP-binding protein)
yybL	4174	ABC transporter (permease)
yybO	4169	ABC transporter (permease)
yycB	4159	ABC transporter (permease)
yydI	4125	ABC transporter (ATP-binding protein)
yyzE	4122	phosphotransferase systeme enzyme II

I.3 SENSORS (SIGNAL TRANSDUCTION) 38

cheA	1712	two-component sensor histidine kinase [CheB/CheY] chemotactic signal modulator
citS	830	two-component sensor histidine kinase [CitT]
comP	3255	two-component sensor histidine kinase [ComA] involved in early competence
degS	3646	two-component sensor histidine kinase [DegU] involved in degradative enzyme and competence regulation
kinA	1469	two-component sensor histidine kinase [Spo0F] involved in the initiation of sporulation
kinB	3229	two-component sensor histidine kinase [Spo0F] involved in the initiation of sporulation
kinC	1518	two-component sensor histidine kinase [Spo0A] involved in the initiation of sporulation (phosphorelay-independent)
lytS	2957	two-component sensor histidine kinase [LytT] involved in the rate of autolysis
phoR	2977	two-component sensor histidine kinase [PhoP] involved in phosphate regulation
resE	2416	two-component sensor histidine kinase [ResD] involved in aerobic and anaerobic respiration
ybdK	222	two-component sensor histidine kinase [YbdJ]
ycbA	266	two-component sensor histidine kinase [YcbB]
ycbM	279	two-component sensor histidine kinase [YcbL]
yccG	295	two-component sensor histidine kinase [YccH]
yclK	427	two-component sensor histidine kinase [YclJ]
ydbF	497	two-component sensor histidine kinase [YdbG]
ydfH	587	two-component sensor histidine kinase [YdfI]
yesM	758	two-component sensor histidine kinase [YesN]
yfiJ	903	two-component sensor histidine kinase [YfiK]
yhcY	1008	two-component sensor histidine kinase [YhcZ]
yhjL	1129	sensory transduction pleiotropic regulatory protein
ykoH	1392	two-component sensor histidine kinase [YkoG]
ykrQ	1419	two-component sensor histidine kinase
ykvD	1432	two-component sensor histidine kinase
yocF	2090	two-component sensor histidine kinase [YocG]
yrkQ	2704	two-component sensor histidine kinase [YrkP]
ytrP	3035	two-component sensor histidine kinase
ytsB	3112	two-component sensor histidine kinase [YtsA]
yufL	3236	two-component sensor histidine kinase [YufM]
yvcQ	3566	two-component sensor histidine kinase [YvcP]
yvfT	3497	two-component sensor histidine kinase [YvfU]
yvqB	3385	two-component sensor histidine kinase [YvqA]
yvqE	3395	two-component sensor histidine kinase [YvqC]
yvrG	3407	two-component sensor histidine kinase [YvrH]
ywpD	3741	two-component sensor histidine kinase
yxdK	4071	two-component sensor histidine kinase [YxdJ]
yxjM	3992	two-component sensor histidine kinase [YxjL]
yycG	4153	two-component sensor histidine kinase [YycF]

I.4 MEMBRANE BIOENERGETICS (ELECTRON TRANSPORT CHAIN AND ATP SYNTHASE) 78

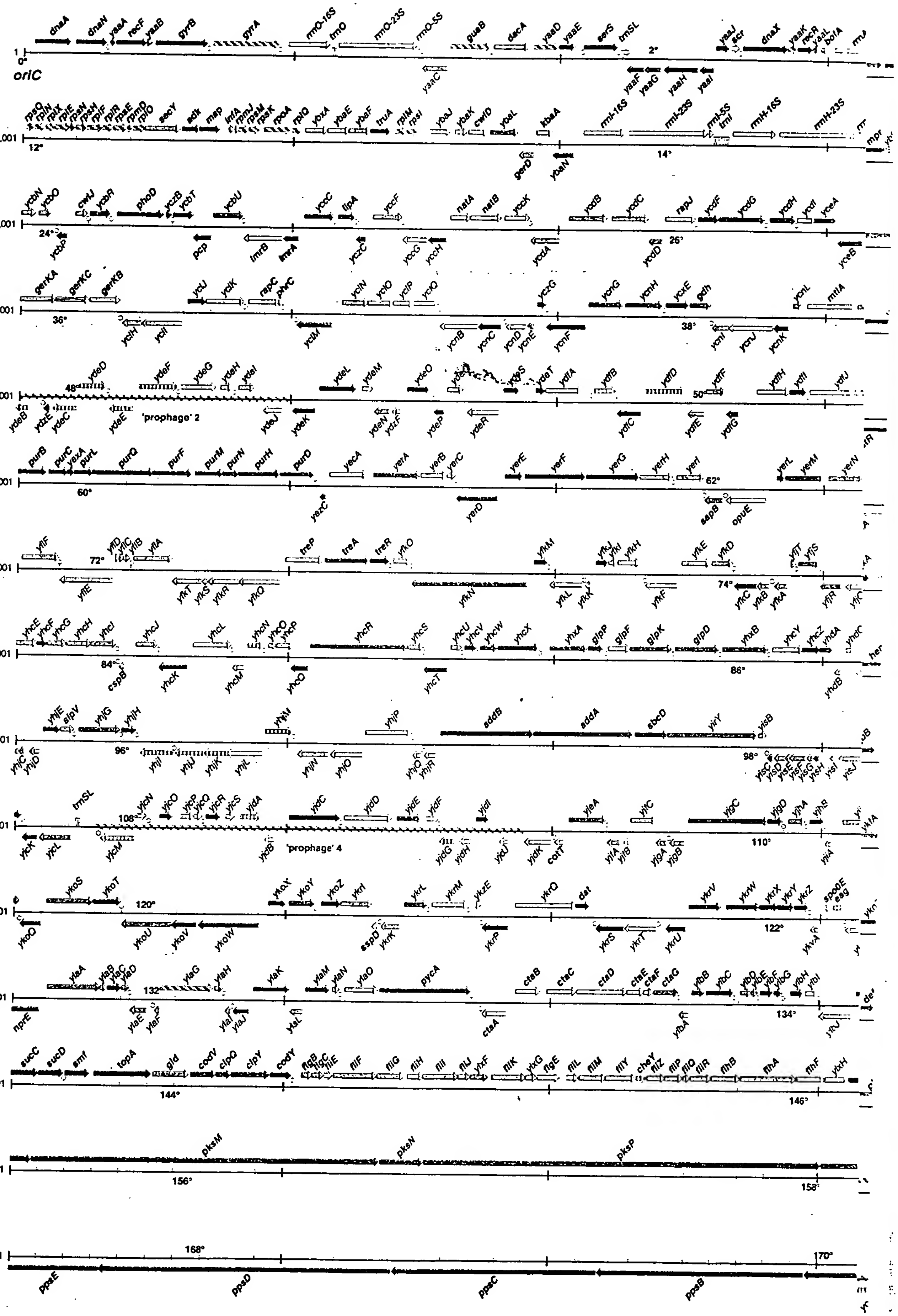
atpA	3784	ATP synthase (subunit α)
atpB	3787	ATP synthase (subunit a)
atpC	3781	ATP synthase (subunit ε)
atpD	3782	ATP synthase (subunit β)
atpE	3786	ATP synthase (subunit c)
atpF	3786	ATP synthase (subunit b)
atpG	3783	ATP synthase (subunit γ)
atpH	3785	ATP synthase (subunit δ)
atpI	3787	ATP synthase (subunit i)
cccA	2599	cytochrome c_{550}
cccB	3625	cytochrome c_{551}
ccdA	1922	required for a late step of cytochrome *c* synthesis
ctaA	1558	cytochrome caa_3 oxidase (required for biosynthesis)
ctaB	1559	cytochrome caa_3 oxidase (assembly factor)
ctaC	1560	cytochrome caa_3 oxidase (subunit II)
ctaD	1561	cytochrome caa_3 oxidase (subunit I)
ctaE	1563	cytochrome caa_3 oxidase (subunit III)
ctaF	1563	cytochrome caa_3 oxidase (subunit IV)
cydA	3978	cytochrome *bd* ubiquinol oxidase (subunit I)
cydB	3977	cytochrome *bd* ubiquinol oxidase (subunit II)
etfA	2915	electron transfer flavoprotein (α subunit)
etfB	2916	electron transfer flavoprotein (β subunit)
fer	2409	ferredoxin
hmp	1372	flavohemoglobin
narG	3829	nitrate reductase (α subunit)
narH	3825	nitrate reductase (β subunit)
narI	3823	nitrate reductase (γ subunit)
narJ	3824	nitrate reductase (protein J)
ndhF	205	NADH dehydrogenase (subunit 5)
qcrA	2364	menaquinol:cytochrome *c* oxidoreductase (iron-sulphur subunit)
qcrB	2364	menaquinol:cytochrome *c* oxidoreductase (cytochrome *b* subunit)
qcrC	2363	menaquinol:cytochrome *c* oxidoreductase (cytochrome *b/c* subunit)
qoxA	3917	cytochrome aa_3 quinol oxidase (subunit II)
qoxB	3916	cytochrome aa_3 quinol oxidase (subunit I)
qoxC	3914	cytochrome aa_3 quinol oxidase (subunit III)
qoxD	3913	cytochrome aa_3 quinol oxidase (subunit IV)
resA	2421	essential protein similar to cytochrome *c* biogenesis protein
resB	2420	essential protein similar to cytochrome *c* biogenesis protein
resC	2418	essential protein similar to cytochrome *c* biogenesis protein
tlp	1930	thioredoxin-like protein
trxA	2912	thioredoxin
trxB	3573	thioredoxin reductase
ycgT	352	thioredoxin reductase
ycnD	439	NADPH-flavin oxidoreductase
ydbP	508	thioredoxin
ydeQ	576	NAD(P)H oxidoreductase
ydfQ	598	thioredoxin
ydgI	613	NADH dehydrogenase
yfkO	854	NAD(P)H-flavin oxidoreductase
yfmJ	818	quinone oxidoreductase
yjdK	1280	cytochrome *c* oxidase assembly factor
yjlD	1299	NADH dehydrogenase
ykuN	1486	flavodoxin
ykuP	1488	sulfite reductase
ykuU	1492	2-cys peroxiredoxin
ykvV	1450	thioredoxin
yneN	1929	thiol:disulfide interchange protein
yojN	2114	nitric-oxide reductase
yolI	2267	thioredoxin
yosR	2159	thioredoxin
ypdA	2401	thioredoxin reductase
yqiG	2516	NADH-dependent flavin oxidoreductase
yqjM	2475	NADH-dependent flavin oxidoreductase
yrkL	2708	NAD(P)H oxidoreductase
ythA	3139	cytochrome *d* oxidase subunit
ytpP	3054	thioredoxin H1
ytrC	3117	cytochrome *c* oxidase subunit
ytrD	3116	cytochrome *c* oxidase subunit
yufD	3249	NADH dehydrogenase (ubiquinone)
yufT	3246	NADH dehydrogenase
yumB	3300	NADH dehydrogenase
yumC	3301	thioredoxin reductase
yusE	3364	thioredoxin
yutJ	3308	NADH dehydrogenase
yvaB	3445	NAD(P)H dehydrogenase (quinone)
ywcG	3911	NADPH-flavin oxidoreductase
ywhN	3840	ubiquinol-cytochrome *c* reductase
ywrO	3708	NAD(P)H oxidoreductase

I.5 MOBILITY AND CHEMOTAXIS 55

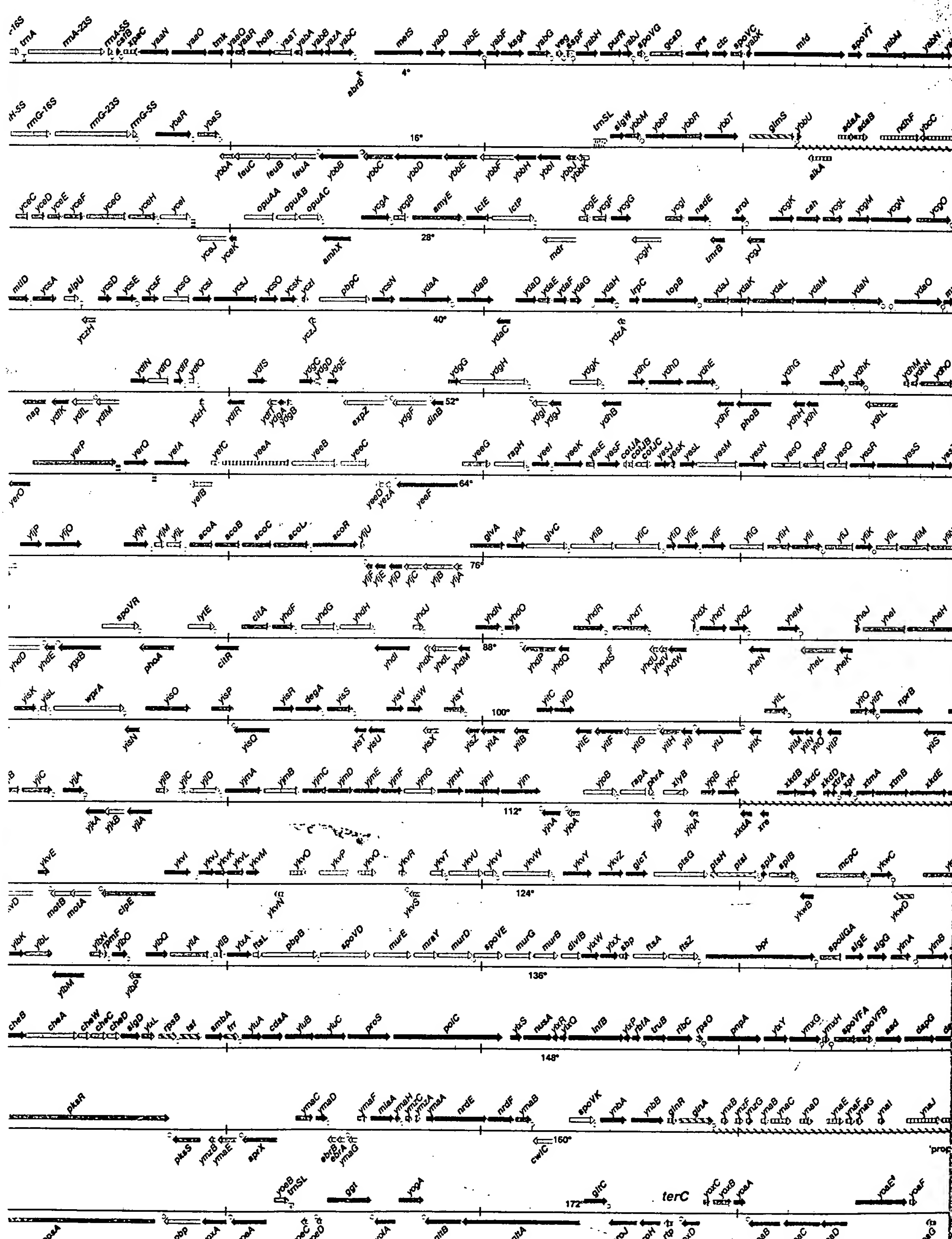
cheC	1715	inhibition of CheR-mediated methylation of methyl-accepting chemotaxis proteins
cheD	1715	required for methylation of methyl-accepting chemotaxis proteins by CheR
cheR	2380	methyl-accepting chemotaxis proteins methyltransferase
cheV	1473	modulation of CheA activity in response to attractants (CheW and CheY similar domains)
cheW	1714	modulation of CheA activity in response to attractants
flgB	1691	flagellar basal-body rod protein
flgC	1691	flagellar basal-body rod protein

Table 1. Functional classification of the *Bacillus subtilis* protein-coding genes.

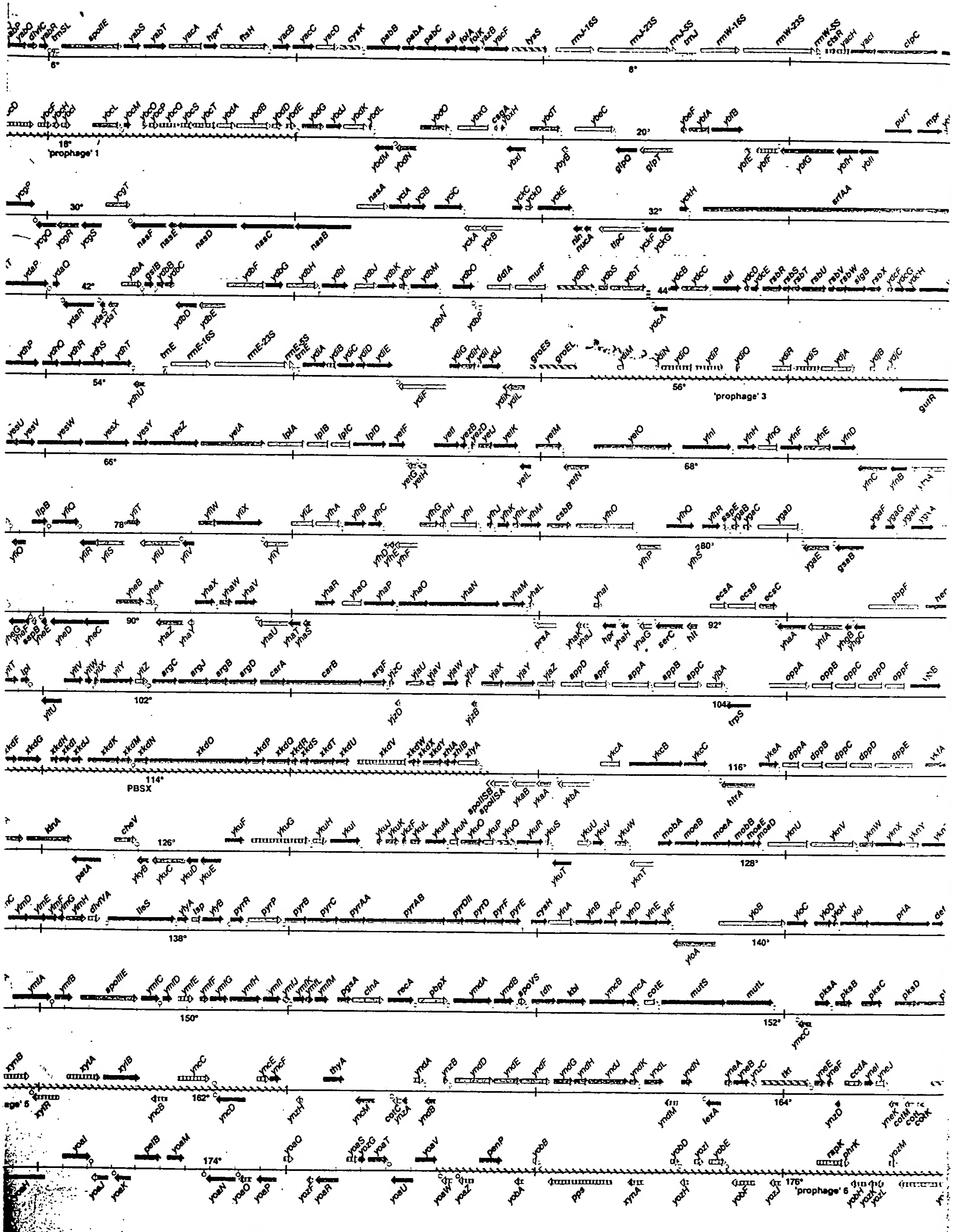
I		CELL ENVELOPE AND CELLULAR PROCESSES 866
1.1		CELL WALL 93
cwlA	2665	*N*-acetylmuramoyl-L-alanine amidase (minor autolysin)
cwlC	1873	*N*-acetylmuramoyl-L-alanine amidase (sporulation mother cell wall)
cwlD	157	*N*-acetylmuramoyl-L-alanine amidase (germination)
cwlJ	282	cell wall hydrolase (sporulation)
dacA	18	penicillin-binding protein 5 (D-alanyl-D-alanine carboxypeptidase) (peptidoglycan biosynthesis)
dacB	2424	penicillin-binding protein 5* (D-alanyl-D-alanine carboxypeptidase) (peptidoglycan biosynthesis) (spore cortex)
dacF	2445	penicillin-binding protein (D-alanyl-D-alanine carboxypeptidase) (peptidoglycan biosynthesis)
ddlA	508	D-alanyl-D-alanine ligase A (peptidoglycan biosynthesis)
dltA	3951	D-alanyl-D-alanine carrier protein ligase (lipoteichoic acid biosynthesis)
dltB	3953	D-alanine transfer from Dcp to undecaprenol-phosphate (lipoteichoic acid biosynthesis)
dltC	3954	D-alanine carrier protein (lipoteichoic acid biosynthesis)
dltD	3954	D-alanine transfer from undecaprenol-phosphate to the poly(glycerophosphate) chain (lipoteichoic acid biosynthesis)
dltE	3955	involved in lipoteichoic acid biosynthesis
gcaD	56	UDP-*N*-acetylglucosamine pyrophosphorylase (peptidoglycan and lipopolysaccharide biosynthesis)
ggaA	3670	galactosamine-containing minor teichoic acid biosynthesis
ggaB	3669	galactosamine-containing minor teichoic acid biosynthesis
gtaB	3665	UTP-glucose-1-phosphate uridylyltransferase
lytB	3662	modifier protein of major autolysin LytC (CWBP76)
lytC	3660	*N*-acetylmuramoyl-L-alanine amidase (major autolysin) (CWBP49)
lytD	3687	*N*-acetylglucosaminidase (major autolysin) (CWBP90)
lytE	1018	cell wall lytic activity (CWBP33)
mbl	3747	MreB-like protein
mraY	1587	phospho-*N*-acetylmuramoyl-pentapeptide transferase (peptidoglycan biosynthesis)
mreB	2861	cell-shape determining protein
mreBH	1517	cell-shape determining protein
mreC	2860	cell-shape determining protein
mreD	2859	cell-shape determining protein
murA	3778	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase (peptidoglycan biosynthesis)
murB	1592	UDP-*N*-acetylenolpyruvoylglucosamine reductase (peptidoglycan biosynthesis)
murC	3049	UDP-*N*-acetylmuramate-alanine ligase (peptidoglycan biosynthesis)
murD	1588	UDP-*N*-acetylmuramoylalanine-D-glutamate ligase (peptidoglycan biosynthesis)
murE	1586	UDP-*N*-acetylmuramoylananine-D-glutamate-2,6-diaminopimelate ligase (peptidoglycan biosynthesis)
murF	509	UDP-*N*-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl-D-alanyl ligase (peptidoglycan biosynthesis)
murG	1591	UDP-*N*-acetylglucosamine-*N*-acetylmuramyl-(pentapeptide)pyrophosphoryl-undecaprenol *N*-acetylglucosamine transferase (peptidoglycan biosynthesis)
murZ	3806	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase (peptidoglycan biosynthesis)
pbp	1999	penicillin-binding protein (peptidoglycan biosynthesis)
pbpA	2583	penicillin-binding protein 2A (peptidoglycan biosynthesis) (spore outgrowth)
pbpB	1581	penicillin-binding protein 2B (peptidoglycan biosynthesis) (cell-division septum)
pbpC	463	penicillin-binding protein 3 (peptidoglycan biosynthesis)
pbpD	3233	penicillin-binding protein 4 (peptidoglycan biosynthesis)
pbpE	3535	penicillin-binding protein 4* (peptidoglycan biosynthesis) (spore cortex)
pbpF	1083	penicillin-binding protein 1A (peptidoglycan biosynthesis) (germination)
pbpX	1765	penicillin-binding protein (peptidoglycan biosynthesis)
ponA	2341	penicillin-binding proteins 1A/1B (peptidoglycan biosynthesis)
racE	2903	glutamate racemase (peptidoglycan biosynthesis)
spoVD	1584	penicillin-binding protein (peptidoglycan biosynthesis) (spore cortex)
tagA	3680	involved in polyglycerol phosphate teichoic acid biosynthesis
tagB	3681	involved in polyglycerol phosphate teichoic acid biosynthesis
tagC	3682	involved in polyglycerol phosphate teichoic acid biosynthesis
tagD	3680	glycerol-3-phosphate cytidylyltransferase (teichoic acid biosynthesis)
tagE	3679	UDP-glucose:polyglycerol phosphate glucosyltransferase (teichoic acid biosynthesis)
tagF	3677	CDP-glycerol:polyglycerol phosphate glycerophosphotransferase (teichoic acid biosynthesis)
tagG	3675	teichoic acid translocation (permease)
tagH	3674	teichoic acid translocation (ATP-binding protein)
tagO	3649	teichoic acid linkage unit synthesis
tuaA	3658	biosynthesis of teichuronic acid
tuaB	3657	biosynthesis of teichuronic acid
tuaC	3656	biosynthesis of teichuronic acid
tuaD	3655	biosynthesis of teichuronic acid (UDP-glucose 6-dehydrogenase)
tuaE	3653	biosynthesis of teichuronic acid
tuaF	3652	biosynthesis of teichuronic acid
tuaG	3651	biosynthesis of teichuronic acid
tuaH	3650	biosynthesis of teichuronic acid
wapA	4029	cell wall-associated protein precursor (CWBP200, 105, 62)
wprA	1153	cell wall-associated protein precursor (CWBP23 and serine protease CWBP52)
xlyA	1347	*N*-acetylmuramoyl-L-alanine amidase (PBSX prophage-mediated lysis)
xlyB	1317	*N*-acetylmuramoyl-L-alanine amidase (PBSX prophage-mediated lysis)
yfnG	799	CDP-glucose 4,6-dehydratase
yhdD	1013	cell wall-binding protein
ykuA	1467	penicillin-binding protein
ylbI	1569	lipopolysaccharide core biosynthesis
ymaG	1865	cell wall protein
yngB	1946	UTP-glucose-1-phosphate uridylyltransferase
yocH	2093	cell wall-binding protein
yodJ	2135	D-alanyl-D-alanine carboxypeptidase
yojL	2116	cell wall-binding protein
yomC	2263	*N*-acetylmuramoyl-L-alanine amidase
ypdQ	2310	cell wall enzyme
ypfP	2306	cell wall synthesis
ypjH	2357	lipopolysaccharide biosynthesis-related protein
yqeE	2649	*N*-acetylmuramoyl-L-alanine amidase
yqfY	2588	peptidoglycan acetylation
yqiI	2515	*N*-acetylmuramoyl-L-alanine amidase
yrhL	2771	acyltransferase
yrrR	2791	penicillin-binding protein
yrvJ	2818	*N*-acetylmuramoyl-L-alanine amidase
ytcC	3157	lipopolysaccharide *N*-acetylglucosaminyltransferase
ytkC	3135	autolytic amidase
ytxN	3161	lipopolysaccharide *N*-acetylglucosaminyltransferase
yubE	3191	*N*-acetylmuramoyl-L-alanine amidase
yvcE	3575	cell wall-binding protein
ywhE	3849	penicillin-binding protein
ywtD	3697	murein hydrolase
1.2		TRANSPORT/BINDING PROTEINS AND LIPOPROTEINS 381
aapA	2766	amino acid permease
alsT	1938	amino acid carrier protein
amyC	3099	maltose transport protein
amyD	3098	sugar transport
appA	1213	oligopeptide ABC transporter (oligopeptide-binding protein)
appB	1215	oligopeptide ABC transporter (permease)
appC	1216	oligopeptide ABC transporter (permease)
appD	1211	oligopeptide ABC transporter (ATP-binding protein)
appF	1212	oligopeptide ABC transporter (ATP-binding protein)
araE	3485	L-arabinose transport (permease)
araN	2942	L-arabinose transport (sugar-binding protein)
araP	2941	L-arabinose transport (integral membrane protein)
araQ	2940	L-arabinose transport (integral membrane protein)
azlC	2729	branched-chain amino acid transport
azlD	2728	branched-chain amino acid transport
bglP	4034	phosphotransferase system (PTS) β-glucoside-specific enzyme IIABC component
blt	2716	multidrug-efflux transporter
bmr	2494	multidrug-efflux transporter
braB	3027	branched-chain amino acid transporter
brnQ	2728	branched-chain amino acid transporter
citM	834	secondary transporter of the Mg^{2+}/citrate complex
csbX	2838	α-ketoglutarate permease
cydC	3976	ABC transporter required for expression of cytochrome *bd* (ATP-binding protein)
cydD	3974	ABC transporter required for expression of cytochrome *bd* (ATP-binding protein)
czcD	2724	cation-efflux system membrane protein
dppA	1360	dipeptide ABC transporter (sporulation)
dppB	1361	dipeptide ABC transporter (permease) (sporulation)
dppC	1362	dipeptide ABC transporter (permease) (sporulation)
dppD	1363	dipeptide ABC transporter (ATP-binding protein) (sporulation)
dppE	1364	dipeptide ABC transporter (dipeptide-binding protein) (sporulation)
ebrA	1865	multidrug resistance protein
ebrB	1864	multidrug resistance protein
ecsA	1077	ABC transporter (ATP-binding protein)
ecsB	1078	ABC transporter (membrane protein)
expZ	606	ATP-binding transport protein
feuA	183	iron-uptake system (binding protein)
feuB	182	iron-uptake system (integral membrane protein)
feuC	181	iron-uptake system (integral membrane protein)
fhuB	3417	ferrichrome ABC transporter (permease)
fhuC	3415	ferrichrome ABC transporter (ATP-binding protein)
fhuD	3418	ferrichrome ABC transporter (ferrichrome-binding protein)
fhuG	3416	ferrichrome ABC transporter (permease)
fruA	1509	phosphotransferase system (PTS) fructose-specific enzyme IIBC component
gabP	686	γ-aminobutyrate permease
glnH	2802	glutamine ABC transporter (glutamine-binding)
glnM	2803	glutamine ABC transporter (membrane protein)
glnP	2804	glutamine ABC transporter (membrane protein)
glnQ	2802	glutamine ABC transporter (ATP-binding protein)
glpF	1002	glycerol uptake facilitator
glpT	235	glycerol-3-phosphate permease
gltP	255	H^+/glutamate symport protein
gltT	1097	H^+/Na^+-glutamate symport protein
glvC	892	phosphotransferase system (PTS) arbutin-like enzyme IIBC component
gntP	4115	gluconate permease (gluconate utilization)
hisP	3004	histidine transport protein (ATP-binding protein)
hutM	4046	histidine permease
iolF	4077	inositol transport protein
kdgT	2322	2-keto-3-deoxygluconate permease (pectin utilization)
lctP	330	L-lactate permease
levD	2762	phosphotransferase system (PTS) fructose-specific enzyme IIA component
levE	2762	phosphotransferase system (PTS) fructose-specific enzyme IIB component
levF	2761	phosphotransferase system (PTS) fructose-specific enzyme IIC component
levG	2760	phosphotransferase system (PTS) fructose-specific enzyme IID component
licA	3959	phosphotransferase system (PTS) lichenan-specific enzyme IIA component
licB	3961	phosphotransferase system (PTS) lichenan-specific enzyme IIB component
licC	3960	phosphotransferase system (PTS) lichenan-specific enzyme IIC component
lmrB	290	lincomycin-resistance protein
lplA	779	lipoprotein
lplB	781	transmembrane lipoprotein
lplC	782	transmembrane lipoprotein
mdr	334	multidrug-efflux transporter (puromycin, norfloxacin, tosufloxacin)
msmE	3097	multiple sugar-binding protein
msmX	3984	multiple sugar-binding transport ATP-binding protein
mtlA	449	phosphotransferase system (PTS) mannitol-specific enzyme IIABC component
narK	3833	nitrite extrusion protein
nasA	363	nitrate transporter
natA	296	Na^+ ABC transporter (extrusion) (ATP-binding protein)
natB	297	Na^+ ABC transporter (extrusion) (membrane protein)
nrgA	3756	ammonium transporter
nupC	4050	pyrimidine-nucleoside transport protein
oppA	1219	oligopeptide ABC transporter (binding protein) (initiation of sporulation, competence development)
oppB	1221	oligopeptide ABC transporter (permease) (initiation of sporulation, competence development)
oppC	1222	oligopeptide ABC transporter (permease) (initiation of sporulation, competence development)
oppD	1223	oligopeptide ABC transporter (ATP-binding protein) (initiation of sporulation, competence development)
oppF	1224	oligopeptide ABC transporter (ATP-binding protein) (initiation of sporulation, competence development)
opuAA	321	glycine betaine ABC transporter (ATP-binding protein) (osmoprotection)
opuAB	322	glycine betaine ABC transporter (permease) (osmoprotection)
opuAC	323	glycine betaine ABC transporter (glycine betaine-binding protein) (osmoprotection)
opuBA	3462	choline ABC transporter (ATP-binding protein) (osmoprotection)
opuBB	3461	choline ABC transporter (membrane protein) (osmoprotection)
opuBC	3460	choline ABC transporter (choline-binding protein) (osmoprotection)
opuBD	3460	choline ABC transporter (membrane protein) (osmoprotection)
opuCA	3470	glycine betaine/carnitine/choline ABC transporter (ATP-binding protein) (osmoprotection)
opuCB	3469	glycine betaine/carnitine/choline ABC transporter (membrane protein) (osmoprotection)
opuCC	3468	glycine betaine/carnitine/choline ABC transporter (osmoprotectant-binding protein) (osmoprotection)
opuCD	3467	glycine betaine/carnitine/choline ABC transporter (membrane protein) (osmoprotection)
opuD	3076	glycine betaine transporter (osmoprotection)
opuE	728	proline transporter (osmoprotection)
pbuX	2319	xanthine permease
ptsG	1457	phosphotransferase system (PTS) glucose-specific enzyme IIABC component
ptsI	1459	phosphotransferase system (PTS) enzyme I (general energy coupling protein of the PTS)
pyrP	1618	uracil permease (pyrimidine biosynthesis)
rbsA	3703	ribose ABC transporter (ATP-binding protein)
rbsB	3705	ribose ABC transporter (ribose-binding protein)
rbsC	3704	ribose ABC transporter (permease)
rbsD	3702	ribose ABC transporter (membrane protein)
rocC	3876	amino acid permease (arginine and ornithine utilization)
rocE	4143	amino acid permease (arginine and ornithine utilization)
sacP	3904	phosphotransferase system (PTS) sucrose-specific enzyme IIBC component
slp	1533	small peptidoglycan-associated lipoprotein
sunT	2269	sublancin 168 lantibiotic transporter
tetB	4188	tetracycline resistance protein
treP	850	phosphotransferase system (PTS) trehalose-specific enzyme IIBC component
trkA	2723	potassium uptake
yabM	65	amino acid transporter
ybaE	151	ABC transporter (ATP-binding protein)
ybbF	191	sucrose phosphotransferase enzyme II
ybcL	212	chloramphenicol resistance protein
ybdA	217	ABC transporter (binding protein)
ybdB	218	ABC transporter (permease)
ybeC	231	amino acid transporter
ybfS	257	phosphotransferase system enzyme II
ybgF	262	histidine permease
ybgH	264	sodium/proton-dependent alanine transporter
ybxA	150	ABC transporter (ATP-binding protein)
ybxG	227	amino acid permease
ycbE	270	glucarate transporter
ycbK	277	efflux system
ycbN	280	ABC transporter (ATP-binding protein)
yccK	298	ion channel
ycdI	309	ABC transporter (ATP-binding protein)
yceI	317	transporter
yceJ	320	multidrug-efflux transporter
ycgH	337	amino acid transporter
ycgO	347	proline permease
yckA	368	amino acid ABC transporter (permease)
yckB	368	amino acid ABC transporter (binding protein)
yckI	410	glutamine ABC transporter (ATP-binding protein)
yckJ	410	glutamine ABC transporter (permease)
yckK	411	glutamine ABC transporter (glutamine-binding protein)
yclF	417	di-tripeptide ABC transporter (membrane protein)
yclH	424	ABC transporter (permease)
yclI	426	transporter
yclN	432	ferrichrome ABC transporter (permease)
yclO	433	ferrichrome ABC transporter (permease)
yclP	434	ferrichrome ABC transporter (ATP-binding protein)
yclQ	435	ferrichrome ABC transporter (binding protein)
ycnB	437	multidrug resistance protein
ycnJ	448	copper export protein
ycsG	457	branched chain amino acids transporter
ydbA	493	ABC transporter (binding protein)
ydbE	497	C4-dicarboxylate binding protein
ydbH	500	C4-dicarboxylate transport protein
ydbJ	502	ABC transporter (ATP-binding protein)
ydeG	566	metabolite transport protein



1 0° oriC
dnaA dnaN yaaA recF yaaB gyrB gyrA rrnO-16S trnO rrnO-23S rrnO-5S guaB dacA yaaD yaaE serS trnSL 2° yaaJ tadA dnaX yaaK recR yaaL bofA
yaaC yaaF yaaG yaaH yaaI
140,001 12° 14°
rpsQ rplN rplX rplE rpsN rpsH rplF rplR rpsE rpmD rplO secY adk map infA rpmJ rpsM rpsK rpoA rplQ ybxA ybaE ybaF truA rplM rpsI ybaJ ybaK cwlD ybaL kbaA rrnI-16S rrnI-23S rrnI-5S trnI rrnH-16S rrnH-23S
gerD ybaN
280,001 24° 26°
ycbN ycbO cwlJ ycbR phoD yczB ycbT ycbU yccC lipA yccF natA natB yccK ycdB ycdC rapJ ycdF ycdG ycdH ycdI yceA
ycbP pcp lmrB lmrA yczC yccG yccH ycdA ycdD yceB
420,001 36° 38°
gerKA gerKC gerKB yclJ yclK rapC phrC yclN yclO yclP yclQ yczG ycnG ycnH ycxE gdh ycnL mtlA
yclH yclI yclM ycnB ycnC ycnD ycnE ycnF ycnI ycnJ ycnK
560,001 48° 50°
ydeD ydeF ydeG ydeH ydeI ydeL ydeM ydeO ydeS ydeT ydfA ydfB ydfD ydfF ydfH ydfI ydfJ
ydeB ydzE ydeC ydeE 'prophage' 2 ydeJ ydeK ydeN ydzF ydeP ydeR ydfC ydfE ydfG
700,001 60° 62°
purB purC yexA purL purQ purF purM purN purH purD yecA yerA yerB yerC yerE yerF yerG yerH yerI yerL yerM yerN
yezC yerD sspB opuE
840,001 72° 74°
yflF yflD yflC yflB yflA treP treA treR yfkO yfkM yfkJ yfkI yfkH yfkE yfkD yfjT yfjS
yflE yflT yflS yflR yflQ yfkN yfkL yfkK yfkF yfkC yfkB yfkA yfjR
980,001 84° 86°
yhcE yhcF yhcG yhcH yhcI yhcJ yhcL yhcN yhcO yhcP yhcR yhcS yhcU yhcV yhcW yhcX yhxA glpP glpF glpK glpD yhxB yhcY yhcZ yhdA
cspB yhcK yhcM yhcQ yhcT yhdB
1,120,001 96° 98°
yhjE sipV yhjG yhjH yhjM yhjP addB addA sbcD yirY yisB
yhjC yhjD yhjI yhjJ yhjK yhjL yhjN yhjO yhjQ yhjR yisC yisD yisE yisF yisG yisH yisI yisJ
1,260,001 108° 110°
trnSL yjcN yjcO yjcP yjcQ yjcR yjcS yjdA yjdC yjdD yjdE yjdF yjdI yjeA yjiC yjgC yjgD yjhA yjhB
yjcK yjcL yjcM yjdB 'prophage' 4 yjdG yjdH yjdJ yjdK cotT yjiA yjiB yjgA yjgB yjjA
1,400,001 120° 122°
ykoS ykoT ykoX ykoY ykoZ ykrI ykrL ykrM ykzE ykrQ dat ykrV ykrW ykrX ykrY ykrZ spo0E
ykoQ ykoU ykoV ykoW aspD ykrK ykrP ykrS ykrT ykrU ykvA
1,540,001 132° 134°
ylaA ylaB ylaC ylaD ylaG ylaH ylaK ylaM ylaN ylaO pycA ctaB ctaC ctaD ctaE ctaF ctaG ylbB ylbC ylbD ylbE ylbF ylbG ylbH ylbI
nprE ylaE ylaF ylaI ylaJ ylaL ctaA ylbA
1,680,001 144° 146°
sucC sucD smf topA gid codV clpQ clpY codY flgB flgC fliE fliF fliG fliH fliI fliJ ylxF fliK ylxG flgE fliL fliM fliY cheY fliZ fliP fliQ fliR flhB flhA flhF ylxH
1,820,001 156° 158°
pksM pksN pksP
1,960,001 168° 170°
ppsE ppsD ppsC ppsB

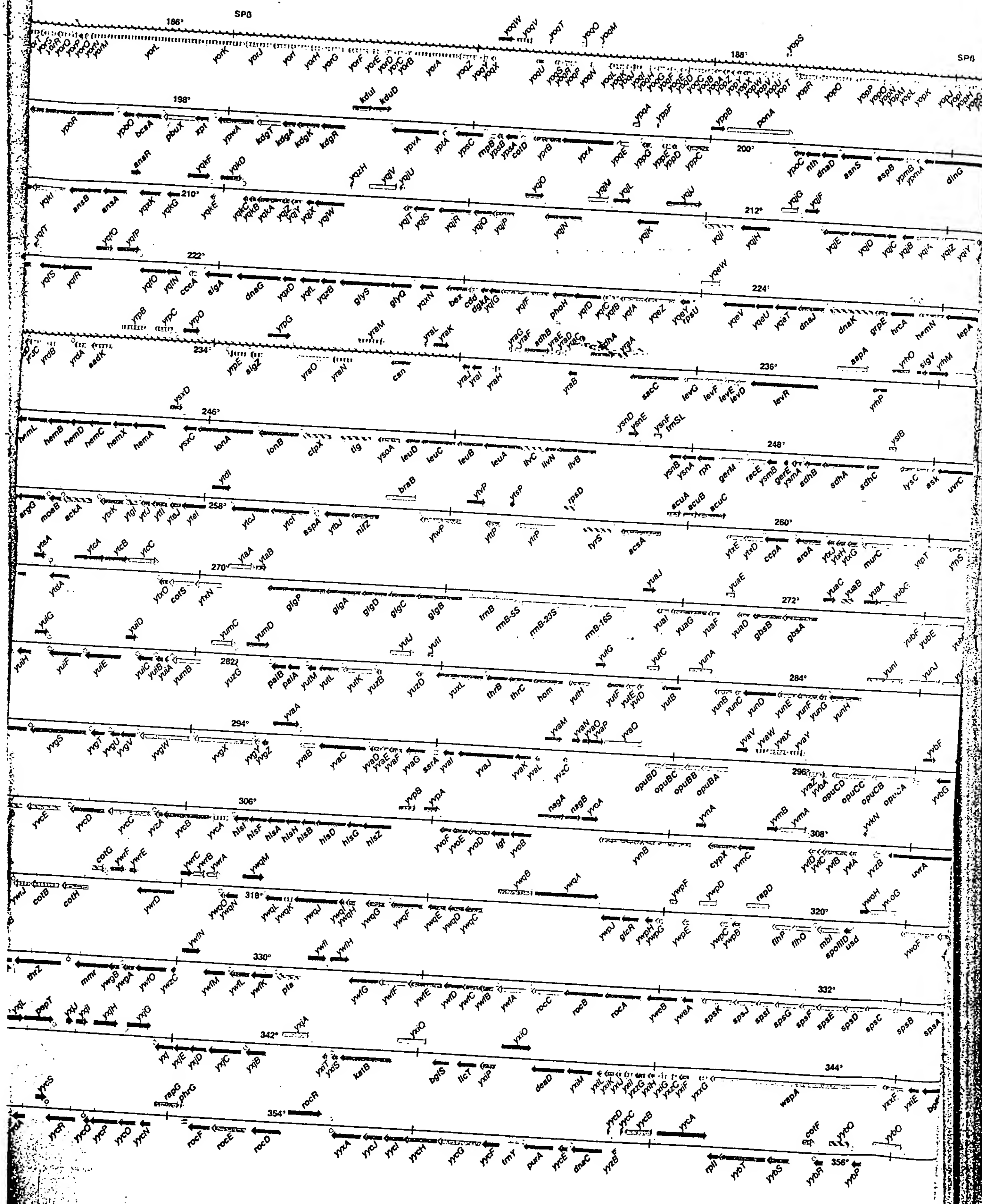


trnA rrnA-23S rrnA-5S xpaC yaaN yaaO tmk holB metS yabD yabE yabF ksgA yabG veg purR spoVG gcaD prs ctc spoVC mfd spoVT yabM yabN
abrB
4°
rrnG-16S rrnG-23S rrnG-5S ybaR ybaS trnSL sigW ybbP ybbR ybbT glmS ybbU adaA adaB ndhF ybcC
16°
ybbA feuC feuB feuA ybbB ybbC ybbD ybbE ybbF ybbH ybbI ybbJ ybbK alkA
yceC yceD yceE yceF yceG yceH yceI opuAA opuAB opuAC ycgA ycgB amyE lctE lctP ycgE ycgF ycgG ycgI nadE aroI ycgK cah ycgL ycgM ycgN ycgO
28°
yceJ yceK amhX mdr ycgH tmrB ycgJ
mtlD ycsA sipU ycsD ycsE ycsF ycsG ycsI ycsJ ycsO ycsK yczI pbpC ycsN ydaA ydaB ydaD ydaE ydaF ydaG ydaH lrpC topB ydaJ ydaK ydaL ydaM ydaN ydaO
40°
yczH yczJ ydaC ydzA
ydfN ydfO ydfP ydfQ ydfS ydgC ydgD ydgE ydgG ydgH ydgK ydhC ydhD ydhE ydhG ydhJ ydhK ydhM ydhN ydhO
nap ydfK ydfL ydfM ydzH ydfR ydfT ydgA ydgB expZ ydgF dinB 52° ydgI ydgJ ydhB ydhF phoB ydhH ydhI ydhL
yerP yerQ yetA yefC yeeA yeeB yeeC yeeG rapH yeeI yeeK yesE yesF cotJA cotJB cotJC yesJ yesK yesL yesM yesN yesO yesP yesQ yesR yesS
yerO yetB yeeD yezA yeeF 64°
yfiP yfiO yfiN yfiM yfiL acoA acoB acoC acoL acoR yfjU glvA yfiA glvC yfiB yfiC yfiD yfiE yfiF yfiG yfiH yfiI yfiJ yfiK yfiL yfiM
yfjF yfjE yfjD yfjC yfjB yfjA 76°
spoVR lytE citA yhdF yhdG yhdH yhdJ yhdN yhdO yhcR yhcT yhdX yhdY yhdZ yheM yheJ yheI yheH
yhcD yhdE ygxB phoA citR yhdI yhdK yhdL yhdM 88° yhdP yhdQ yhdS yhdU yhdV yhdW yheN yheL yheK
yisK yisL wprA yisO yisP yisR degA yisS yisV yisW yisY 100° yitC yitD yitL yitQ yitR nprB
yisN yisQ yisT yisU yisX yisZ yitA yitB yitE yitF yitG yitH yitI yitJ yitK yitM yitN yitO yitP yitS
yjiC yjjA yjiB yjiC yjiD yjmA yjmB yjmC yjmD yjmE yjmF yjmG yjmH yjmI yjm yjoB rapA phrA xlyB yjqB yjqC xkdB xkdC xkdD xtrA xpf xtmA xtmB xkdE
yjkA yjkB yjlA 112° yjnA yjoA yjp yjqA xkdA xre
ykvE ykvI ykvJ ykvK ykvL ykvM ykvO ykvP ykvQ ykvR ykvT ykvU ykvV ykvW ykvY ykvZ glcT ptsG ptsH ptsI splA splB mcpC ykwC
motB motA clpE ykvN ykvS 124° ykwB ykwD
ylbK ylbL ylbN rpmF ylbO ylbQ yllA ylxA ftsL pbpB spoVD murE mraY murD spoVE murG murB divIB ylxW ylxX sbp ftsA ftsZ bpr spoIIGA sigE sigG ylmA ylmB
ylbM ylbP 136°
cheB cheA cheW cheC cheD sigD ylxL rpsB tsf smbA frr ylxA cdsA ylxB ylxC proS polC ylxS nusA ylxR ylxQ infB ylxP rbfA truB ribC rpsO pnpA ylxY ymxG ymxH spoVFA spoVFB asd dapG
148°
pksR ymaC ymaD ymaF miaA ymaH ymzC ymzA nrdE nrdF ymaB spoVK ynbA ynbB glnR glnA
pksS ymzB ymaE aprX abrB ebrA ymaG cwlC 160° 'prop
ppsA pbp yoxA yoeA yoeB trnSL yoeC yoeD ggt yotA yogA gltB gltA 172° gltC proJ proH rtp yoxD terC yoxC yoxB yoaA yoaB yoaC yoaD yoaE yoaF yoaG


6°
8°
16°
'prophage' 1
20°
30°
32°
42°
44°
54°
56°
'prophage' 3
66°
68°
78°
80°
90°
92°
102°
104°
114°
PBSX
116°
126°
128°
138°
140°
150°
152°
162°
164°
174°
176°
'prophage' 6

sms yacK yacL yacM yacN gltX cysE cysS yazC yacO yacP sigH rpmG secE nusG rplK rplA rplJ rplL ybxB rpoB rpoC ybxF rpsL rpsG fus tufA ybaC rpsJ rplC rplD rplW rplB rpsS rplV rpsC rplP rpmC
10°
140,000
ybfK pssA ybfM psd ybfN ybfO ybfP ybfQ ybgA ybgB ybgE ycbA ycbB ycbC ycbD ycbE ycbF ycbG ycbH ycbJ yczA ycbK ycbL ycbM
22°
gltP ybfS ybfT ybgF ybgG ybgH ybgJ
280,000
comS srfAB srfAC srfAD ycxA ycxD yclB yclC yclD yclE yclG
34°
ycxB ycxC slp yczE yckI yckJ yckK yclA yclF yczF
420,000
dcl ydcK trnS sacV ydcO ydcP ydcQ ydcR ydcS ydcT yddA yddB yddC yddD yddE yddF yddG yddH yddI yddJ rapI phrI yddM lrpA yddQ yddR yddS yddT ydeA cspC
46°
'prophage' 2
ydcL ydcM ydcN yddK yddN lrpB
560,000
gutB ydjD ydjE ydjF ydjG ydjH ydjI ydjK ydjM ydjN yeaB yeaC yeaD yebA guaA yebB yebC yebD yebE yebF yebG purE purK
58°
ydjJ ydjL ydjO ydjP yeaA cotA gabP
700,000
yfmT yfmS yfmR yfmQ yfmP yfmO yfmN yfmL yfmK yfmH yfmG yfmA yflT pel yflS citS citT yflP citM yflN yflM yflK
70°
yfmM yfmJ yfmI yfmF yfmE yfmD yfmC yfmB yflL yflJ yflI yflH yflG
840,000
rrnD-16S rrnD-23S rrnD-5S trnD ygzA ygaJ thiA senS ygaL ygbA ygaM ygcA ygaN yhzA trnSL yhbA yhbB cspR yhbD yhbE yhbF prkA yhbH yhbI yhbJ yhcA yhcB yhcC yhcD
82°
ygaI ygaK katA ygaO yhzB
980,000
hemH hemY yhgD yhgE yhfC yhfE yhfF yhfI yhfJ yhfK yhfL yhfN yhfO yhfP yhfQ yhxC comK yhjA
94°
yhfB yhfD gltT yhfH yhfM aprE yhfR yhfS yhfT yhfU yhfV yhfW yhzC yhxD yhjB
1,120,000
yjbC yjbD mecA yjbF yjbG yjbL yjbM yjbN yjbO yjbQ tenA tenI yjbR yjbS yjbT yjbU yjbV yjbW yjbX yjcA yjcB yjcC yjcI yjcJ
106°
yjbE yjbH yjbI yjbJ yjbK yjbP cotZ cotY cotX cotW cotV yjcD yjcE yjcF yjcG yjcH
1,260,000
ykfB ykfC ykfD ykhA hmp ykjA ykkA ykkB ykkC ykkD ykkE proB proA yklA ykzA lspU ykoG ykoH ykoI ykoJ ykzD ykoK ykoB ykoL ykoM ykoP
118°
ykgB ykgA ykzH yknA metC lspA ykoC ykoD ykoE ykoF tnrA ykoN
1,400,000
fruR fruB fruA slpT ykqA ykqB abh kinC ykqA ykqB adeC ykrA ykyA pdhA pdhB pdhC pdhD yktA ykzI ykrC yktC ykzC
130°
ykoA ampS ykpC mreBH ykqC ykzG ykrB slp cad yktB yktD
1,540,000
fmt ylqM ylqN ylqO ylqP ylqQ ylqR yloS spoVM yloU yloV yloW ylpA ylpB ylpC plsX fabD fabG acpA rncS smc ftsY ylxM ffh rpsP ylqD ylqE trmD rplS ylqF rnh ylqG ylqH
142°
rpmB ylqB
1,680,000
pksE pksF pksG pksH pksI pksJ pksK pksL
154°
1,820,000
citB yneN tlp yneP yneQ yneT gntB gntA sleT bglC ynfE trnSL yngA yngB yngC
166°
yneR yneS yntC ynfF xynD yngD yngE yngF yngG yngH yngI yngJ ynzE yngK yngL
1,960,000
yobN yobO yobW yocA yocC yocD yocE yocF yocG yozN yocN
179°
2,100,000


SPβ
190°
2,240,000
202°
2,380,000
214°
2,520,000
226°
skin
2,660,000
238°
2,800,000
250°
2,940,000
262°
3,080,000
274°
3,220,000
278°
286°
3,360,000
290°
298°
3,500,000
310°
3,640,000
322°
3,780,000
334°
3,920,000
346°
4,060,000
358°
4,200,000


SPβ
186°
188°
SPβ
198°
200°
210°
212°
222°
224°
234°
236°
246°
248°
258°
260°
270°
272°
282°
284°
294°
296°
306°
308°
318°
320°
330°
332°
342°
344°
354°
356°

184°
cgeA cgeB yodU yotL sspC yosU yosA
deoD yodL yodM yozD yodN yozE yodO yodP argE yodR yodS yodT cgeE cgeD cgeC yodV yolN yolM yolK yolJ yolI yotG yotE yotD yotC yotB yosZ yosX yosW yosV yosT yosS yosR yosQ yosP yosO
196°
yokF yokE yokD yokA yppP ypoP ypIP ypIQ metB ypfP capD ypzA ypdP ypdQ
yokJ yokI yokH yokG yokC yokB yppQ yppP ypnP ypmT ypmS ypmR ypmQ ypmP ilvA ypkP dfrA thyB ypjQ ypjP ypiP yphP ilvD ypgR ypgQ bsaA degR ypeQ ypeP ypcP
208°
ypuF ypzC ypzD ypuA
resE resD resC resB resA ypuL spmB spmA dacB ypuI ypuH ypuG ribT ribH ribA ribB ribG ypuE ypuD sipS ypuC ypuB ppiB lysA spoVAF spoVAE spoVAD spoVAC spoVAB spoVAA sigF spoIIAB spoIIAA dacF pnp drm ripX yqkL spoIIIM yqkK yqkJ
220°
yqzG yqhB yqgY yqgW yqzD yqzC yqgB yqfZ yqfY yqfW
yqzE comGG comGF comGE comGD comGC comGB comGA yqxL yqhA trnSL yqgZ yqgX yqgV yqgU yqgT yqgS glcK yqgQ yqgP yqgO yqgN yqgM yqgL yqgK yqgJ yqgI yqgH yqgG pbpA yqgE sodA yqgC yqgA yqfX yqfV
232°
yqaE yqaB spoIIIC yrkO yrkN yrkM blt bltD yrdQ gltR 'prophage' 7
yqaK yqaJ yqaI yqaH yqaG yqdA yqaF yqaD yqaC yrkS yrkR yrkQ yrkP yrkL yrkK yrkJ yrkI yrkH yrkG yrkF yrkE yrkD yrkC yrkB bltR yrkA yrdR trkA czcD yrdN yrdK brnQ azlD azlC azlB yrdF cypA
244°
yrzF yrzG yrzH yrbD nifS yrxA
ruvB ruvA bofC csbX yrbE yrbC yrbB yrbA nadA nadC nadB pheA pheB obg spo0B rpmA ysxB rplU spoIVFB spoIVFA minD minC mreD mreC mreB ysxA maf spoIIB comC folC valS ysxE spoVID
256°
ytvI ytrI ytqI ytlI
phoR phoP citH citC citZ ytwI ytzA pykA pfk accA ytlI ytsJ dnaE ytqI ytoI ytnM hipO ribR ytnJ ytnI ytmO hisP ytmM ytmL ytmK ytmJ ytmI ytkL ytkK ytzD argH
268°
ytqA ytqB ytoA pckA ytlA ytlB ytlC ytlD ytjA ythA ythB ythC
ytrE ytrD ytrC ytrB ytrA ytzC ytpB ytpA ytnA asnB metK ytmB ytmA ytkD ytkC dps ytkA ytjB ytiB ytiA ytgD ytgC ytgB ytgA ytfD menE menB ytxM menD
278°
yuzC yuzE yukF ald yuxI yukJ
280°
comX comQ degQ yuxH yueK yueJ yueI yueH yueG yueF yuzF yueE yueD yueC yueB yukA yukB yukC yukD yukE yukL yukM dhbF dhbB dhbE dhbC dhbA
290°
yvrD yvrK yvrL fhuD yvsG yvgJ yvgL yvgM yvgO
292°
yvqF yvqG yvqH yvqI yvqJ yvqK yvrA yvrB yvrC yvrE yvrG yvrH yvrI yvrM yvrN yvrO yvrP fhuC fhuG fhuB yvsH yvgK yvgN yvgP yvgQ
302°
sacB yveB yveA yvdT yvdS yvdR yvdO clpP yvdC yvcT
304°
yvdQ yvdP trnQ yvdM yvdL yvdK yvdJ yvdI yvdH yvdG yvdF yvdE yvdD yvdB yvdA yvcS yvcR yvcQ yvcP yvcN crh yvcL yvcK yvcJ yvcI
314°
tagA tagB tagC gerBA gerBB gerBC ywtF ywtE rbsR rbsK rbsD rbsA rbsC rbsB ywsB ywsA alsR
316°
tagF tagE tagD lytD pmi ywtG ywtD ywtC ywtB ywtA ywsC ywrO alsD alsS
326°
ywjE ywjD ywjC ywiC sbo ywiA ywhR ywhQ ywhP ywhO ywhN ywhM rapF phrF ywhH ywhE
328°
ywjB ywjA ywiE narI narJ narH narG ywiD fnr narK argS ywiB ywhL ywhK ywhG ywhF ywhD ywhC
338°
dltE ywaA katX yxlH yxlA yxkJ yxkI yxzE aldY yxkC yxjO yxjN
340°
licH licA licC licB licR yxzF yxlJ yxlG yxlF yxlE yxlD yxlC sigY yxkO cydD cydC cydB cydA yxkH msmX yxkF yxkD galE yxkA
350°
yxbB yxbA yxnB asnH yxaM yxaI yxnA yxaC gntR gntK gntP gntZ ahpC ahpF yydK fbp
352°
yxbC yxaL yxaK yxaJ yxaH yxaG yxaF yxaD yxaB yxaA bglA yyzE yydJ yydI yydH yydG yydF yydD yydC

2,100,001
180
182°
yocR odhB odhA yojO yojN yojM yojL yojK yojJ yojG yojF yojE yojC yojB yojA yodB yodD yodE ctpA
yomT yomS yomR yomQ yomP yomO yomN yomM yomK yomJ yomI yomH yomG yomF yomE yomD yomC yomB yomA yolC
194°
2,240,001
192°
SPβ
yozP yomL yolK yolJ yolI sunT sunA yolF uvrX yolD yolB yolA
fer serA
204°
206°
2,380,001
ndk gerCC gerCB gerCA mtrB mtrA hbs spoIVA yphF yphE gpsA ypjC ypnB yphA ypgA yplD cmk ypfB yplA ypeB sleB ypdC ypdA ypcA ypbH ypbG ypbF ypbE ypbD recQ ypoB ypaA ypzE aroC ypuN
yqhQ yqhP yqhL yqhH yqhG sinI sinR
216°
218°
2,520,001
recN ahrC yqxC yqiE yqiD yqiC yqiB folD yqhZ yqhY accC accB spoIIIAH spoIIIAG spoIIIAF spoIIIAE spoIIIAD spoIIIAC spoIIIAB spoIIIAA yqhV efp yqhT yqhS yqhR yqhO yqhN yqhM yqhK yqhJ yqhI cotN sipW
yqcG yqcF yqdB skin yqaP
228°
230
2,660,001
yqzJ yqxI cwlA yqxH yqxG yqcE yqcD yqcC yqcB yqcA yqbT yqbS yqbR yqbO yqbP yqbO yqbN yqbM yqbL yqbK yqbJ yqbI yqbH yqbG yqbF yqbE yqbD yqbC yqbB yqbA yqaT yqaS yqaR yqaQ yqaO yqaN yqaM
glnQ glnH glnM glnP yrvP yrvN yrvJ spoVB yrzE
240°
242
2,800,001
yrrI yrrD yrrC yrrB yrrA yrvO yrzC yrvM aspS hisS yrzK yrvI relA apt yrvE yrvD yrvC secF yrbF yrzD yrbG yrbF tgt queA
ysdB yscA yscB ysaA ytbD ytbE
252°
254
2,940,001
araP araN araM araL araD araB araA abnA ysdC ysdA rplT rpmI infC ysbB ysbA lytT lytS thrS ytxC ytxB dnaI dnaB ytcG ytcF gapB ytcD ytaG ytaF mutM polA
ytdP msmR msmE amyD amyC melA ytvA ytrA
264°
3,080,001
yteR yteQ yteP ytcQ ytcP ytbQ bioI bioB bioD bioF bioA bioW ytaP ywfF leuS ytvB ytrB ytsD ytsC ytsB ytsA
yuzA yugF patB kinB kapB yuxJ pbpD yuxK yufL yufM yufN yufO yufP yufQ yufR yufT yufU yufV yufD yufC yufB
276°
3,220,001
pgi yugK yugJ yugI yugH yugG yugE kapD yulK yulS yuxO comA comP
yusN yusO yusP yusT yusZ mrgA yvqA yvqB yuxN gerAA gerAB gerAC
288°
3,360,001
yurZ yusA yusB yusC yusD yusE yusF yusG yusH yusI yusJ yusK yusL yusM yusQ yusR yusS yusU yusV yusW yusX yusY yvrB yvtA citG yvqC
yvfP yvfH yvfG slr pnbA
300°
3,500,001
yvfO lacA yvfM yvfL yvfK lacR yvfI sigL yvfF yvfE yvfD yvfC yvfB yvfA yveT yveS yveR yveQ yveP yveO yveN yveM yveL yveK padC yveG yveF racX pbpE
yvyE yvhJ lytR gtaB
312°
3,640,001
comFC comFB comFA yviA degU degS tagO tuaH tuaG tuaF tuaE tuaD tuaC tuaB tuaA lytC lytB lytA yvyH ggaB ggaA tagH
ywlA ywkD ywkC ywjG
324°
3,780,001
atpC atpD atpG atpA atpH atpF atpE atpB atpI upp glyA ywlG ywlF ywlE ywlD ywlC ywlB spoIIR ywkF ywkE prfA ywkB ywkA tdk rpmE rho ywjI murZ ywjH fbaA spo0F ctrA rpoE acdA
ywcB ywcA ywbH ywbG ywbF ywbE ywbC epr sacX sacY ywaE tyrZ ywaD dltA dltB dltC
336°
3,920,001
galK ywcD ywcC ywbO ywbN ywbM ywbL thiC thiK ywbI ywbD ywbB ywbA gspA ywaF ywaC ywaB
yxeE yxeD yxeB iolR iolS yxcE yxcD yxcC yxbG aldX
348°
4,060,001
yxeK yxeJ yxeI yxeH yxeG yxeF yxeC yxeA yxdM yxdL yxdK yxdJ iolJ iolI iolH iolG iolF iolE iolD iolC iolB iolA htpG yxcA yxbF
yyaC yyaB
360°
4,200,001
4,214,810
yyaE yyaD spo0J soj yyaA gidB gidA thdF jag spoIIIJ rnpA rpmH

Table 1. (continuation) Functional classification of the *Bacillus subtilis* protein-coding genes.

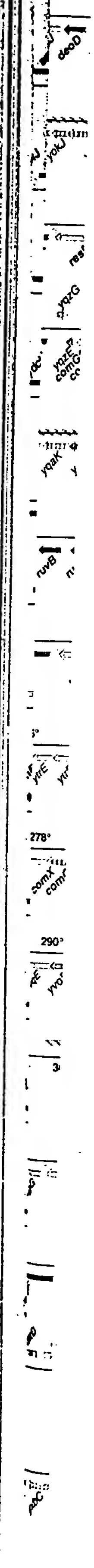
		(valine/isoleucine biosynthesis)
iolA	4083	methylmalonate-semialdehyde dehydrogenase (valine metabolism)
ipi	1189	intracellular proteinase inhibitor
ispA	1386	major intracellular serine protease
kbl	1771	2-amino-3-ketobutyrate CoA ligase
leuA	2893	2-isopropylmalate synthase (leucine biosynthesis)
leuB	2891	3-isopropylmalate dehydrogenase (leucine biosynthesis)
leuC	2890	3-isopropylmalate dehydratase (large subunit) (leucine biosynthesis)
leuD	2889	3-isopropylmalate dehydratase (small subunit) (leucine biosynthesis)
lysA	2437	diaminopimelate decarboxylase (lysine biosynthesis)
lysC	2910	aspartokinase II (α and β subunits) (diaminopimelate/lysine biosynthesis)
metB	2305	homoserine *O*-succinyltransferase (methionine biosynthesis)
metC	1385	cobalamin-independent methionine synthase (methionine biosynthesis)
metK	3128	*S*-adenosylmethionine synthetase
mpr	245	extracellular metalloprotease
nasB	362	assimilatory nitrate reductase (electron transfer subunit)
nasC	360	assimilatory nitrate reductase (catalytic subunit)
nasD	358	assimilatory nitrite reductase (subunit)
nasE	355	assimilatory nitrite reductase (subunit)
nprB	1186	extracellular neutral protease B
nprE	1541	extracellular neutral metalloprotease
nrgB	3757	nitrogen-regulated PII-like protein
patA	1472	aminotransferase
patB	3228	aminotransferase
pepT	3994	peptidase T
pheA	2851	prephenate dehydratase (phenylalanine biosynthesis)
pheB	2852	chorismate mutase (phenylalanine biosynthesis)
proA	1379	γ-glutamyl phosphate reductase (proline biosynthesis)
proB	1378	γ-glutamyl kinase (proline biosynthesis)
proH	2017	involved in proline biosynthesis (salt-inducible)
proJ	2016	glutamate 5-kinase (proline biosynthesis)
racX	3533	amino acid racemase
rocA	3879	pyrroline-5 carboxylate dehydrogenase (arginine and ornithine utilization)
rocB	3878	involved in arginine and ornithine utilization
rocD	4145	ornithine aminotransferase (arginine and ornithine utilization)
rocF	4142	arginase (arginine and ornithine utilization)
serA	2410	phosphoglycerate dehydrogenase (serine biosynthesis)
serC	1076	phosphoserine aminotransferase (serine biosynthesis)
tdh	1770	threonine 3-dehydrogenase (threonine catabolism)
thrB	3313	homoserine kinase (threonine biosynthesis)
thrC	3314	threonine synthase (threonine biosynthesis)
trpA	2372	tryptophan synthase (α subunit) (tryptophan biosynthesis)
trpB	2373	tryptophan synthase (β subunit) (tryptophan biosynthesis)
trpC	2374	indol-3-glycerol phosphate synthase (tryptophan biosynthesis)
trpD	2375	anthranilate phosphoribosyltransferase (tryptophan biosynthesis)
trpE	2377	anthranilate synthase (tryptophan biosynthesis)
trpF	2373	phosphoribosyl anthranilate isomerase (tryptophan biosynthesis)
tyrA	2370	prephenate dehydrogenase (tyrosine biosynthesis)
ureA	3768	urease (γ subunit)
ureB	3768	urease (β subunit)
ureC	3767	urease (α subunit)
vpr	3907	minor extracellular serine protease
yaaO	38	lysine decarboxylase
ybgE	259	branched-chain amino acid aminotransferase
ybgJ	265	glutaminase
yccC	290	asparaginase
ycgM	344	proline oxidase
ycgN	345	1-pyrroline-5-carboxylate dehydrogenase
ydlE	415	prolyl aminopeptidase
yclM	432	homoserine dehydrogenase
ycnG	441	4-aminobutyrate aminotransferase
ycnH	443	succinate-semialdehyde dehydrogenase
ycsA	452	3-isopropylmalate dehydrogenase
ycsJ	459	allophanate hydrolase
yerD	718	glutamate synthase (ferredoxin)
yerM	729	amidase
yhaA	1081	aminoacylase
yhdR	1034	aspartate aminotransferase
yheM	1041	D-alanine aminotransferase
yisK	1152	5-oxo-1,2,5-tricarboxilic-3-penten acid decarboxylase
yisO	1157	asparagine synthase
yisW	1167	opine aminotransferase
yjbG	1231	oligoendopeptidase
yjbR	1243	sarcosine oxidase
yjcI	1258	cystathionine γ-synthase
yjcJ	1259	cystathionine β-lyase
ykeA	1359	pyrroline-5-carboxylate reductase
ykrV	1425	aspartate aminotransferase
ykuQ	1488	tetrahydrodipicolinate succinylase
ykuR	1489	hippurate hydrolase
ylaM	1551	glutaminase
ylmB	1607	acetylornithine deacetylase
yloW	1658	phosphoglycerate dehydrogenase
ylpA	1668	L-serine dehydratase
ymfG	1757	processing protease
ymfH	1758	processing protease
ymxG	1742	processing protease
ynaI	1835	phosphoribosylanthranilate isomerase
yncD	1898	alanine racemase
yobN	2074	L-amino acid oxidase
yodT	2146	adenosylmethionine-8-amino-7-oxononanoate aminotransferase
ypcA	2403	glutamate dehydrogenase
ypwA	2321	carboxypeptidase
yqeI	2644	dihydrodipicolinate reductase
yqhI	2549	aminomethyltransferase
yqhJ	2547	glycine dehydrogenase
yqhK	2546	glycine dehydrogenase
yqhS	2539	3-dehydroquinate dehydratase
yqiT	2503	leucine dehydrogenase
yqiE	2486	tripeptidase
yqiN	2475	amino acid degradation
yqiO	2472	pyrroline-5-carboxylate reductase
yqiR	2470	D-serine dehydratase
yrbE	2839	opine catabolism
yrhA	2786	cysteine synthase
yrhB	2785	cystathionine γ-synthase
yrhP	2768	dihydrodipicolinate reductase
yrpC	2738	glutamate racemase
yrrN	2794	protease
yrrO	2793	protease
ysnE	2897	acetyltransferase
ytfD	3146	*N*-acylamino acid racemase
ytkP	3066	cysteine synthase
yubC	3193	cysteine dioxygenase
yugH	3226	aspartate aminotransferase
yurG	3341	aspartate aminotransferase
yurH	3343	*N*-carbamyl-L-amino acid amidohydrolase
yurL	3347	opine catabolism
yurP	3351	opine catabolism
yurR	3353	opine catabolism
yurT	3354	methylglyoxalase
yusH	3366	glycine cleavage system protein H
yusM	3373	proline dehydrogenase
yusX	3381	oligoendopeptidase
yutL	3306	diaminopimelate epimerase
yuxL	3312	acylaminoacyl-peptidase
yvaK	3454	carboxylesterase
yvfD	3516	serine *O*-acetyltransferase
yvjB	3623	carboxy-terminal processing protease
ywaA	3956	branched-chain amino acid aminotransferase
ywaD	3947	aminopeptidase
yweB	3881	glutamate dehydrogenase
ywfG	3868	aspartate aminotransferase
ywhF	3849	spermidine synthase
ywhG	3848	agmatinase
ywrD	3720	γ-glutamyltransferase
yxeP	4057	aminoacylase

II.3 METABOLISM OF NUCLEOTIDES AND NUCLEIC ACIDS 83

adeC	1521	adenine deaminase
adk	146	adenylate kinase
apt	2823	adenine phosphoribosyltransferase
cdd	2611	cytidine/deoxycytidine deaminase
cmk	2396	cytidylate kinase
ctrA	3811	CTP synthetase (pyrimidine biosynthesis)
deoD	2135	purine nucleoside phosphorylase (purine nucleoside salvage)
dra	4051	deoxyribose-phosphate aldolase (nucleotide/deoxyribonucleotide catabolism)
drm	2448	phosphodeoxyribomutase (purine nucleoside salvage)
guaA	692	GMP synthetase (GMP biosynthesis)
guaB	16	inositol-monophosphate dehydrogenase (GMP biosynthesis)
hipO	3000	hippurate hydrolase
hprT	76	hypoxanthine-guanine phosphoribosyltransferase (purine salvage)
ndk	2381	nucleoside diphosphate kinase
nin	372	inhibitor of the DNA degrading activity of NucA
nrdE	1868	ribonucleoside-diphosphate reductase (major subunit)
nrdF	1870	ribonucleoside-diphosphate reductase (minor subunit)
nucA	372	membrane-associated nuclease
nucB	2652	sporulation-specific extracellular nuclease
pdp	4049	pyrimidine-nucleoside phosphorylase
pnp	2446	purine nucleoside phosphorylase (purine nucleoside salvage)
pnpA	1739	polynucleotide phosphorylase
prs	58	phosphoribosyl pyrophosphate synthetase (nucleotide biosynthesis)
purA	4156	adenylosuccinate synthetase (AMP biosynthesis)
purB	700	adenylosuccinate lyase (purine biosynthesis)
purC	701	phosphoribosylaminoimidazole succinocarboxamide synthetase (purine biosynthesis)
purD	710	phosphoribosylglycinamide synthetase (purine biosynthesis)
purE	698	phosphoribosylaminoimidazole carboxylase I (purine biosynthesis)
purF	705	phosphoribosylpyrophosphate amidotransferase (purine biosynthesis)
purH	708	phosphoribosylaminoimidazole carboxy formyl formyltransferase / inosine-monophosphate cyclohydrolase (purine biosynthesis)
purK	699	phosphoribosylaminoimidazole carboxylase II (purine biosynthesis)
purL	702	phosphoribosylformylglycinamidine synthetase II (purine biosynthesis)
purM	706	phosphoribosylaminoimidazole synthetase (purine biosynthesis)
purN	708	phosphoribosylglycinamide formyltransferase (purine biosynthesis)
purQ	703	phosphoribosylformylglycinamidine synthetase I (purine biosynthesis)
purT	244	phosphoribosylglycinamide formyltransferase 2 (purine biosynthesis)
pyrAA	1622	carbamoyl-phosphate synthetase (glutaminase subunit) (pyrimidine biosynthesis)
pyrAB	1623	carbamoyl-phosphate synthetase (catalytic subunit) (pyrimidine biosynthesis)
pyrB	1620	aspartate carbamoyltransferase (pyrimidine biosynthesis)
pyrC	1621	dihydroorotase (pyrimidine biosynthesis)
pyrD	1627	dihydroorotate dehydrogenase (pyrimidine biosynthesis)
pyrDII	1626	dihydroorotate dehydrogenase (electron transfer subunit) (pyrimidine biosynthesis)
pyrE	1629	orotate phosphoribosyltransferase (pyrimidine biosynthesis)
pyrF	1628	orotidine 5'-phosphate decarboxylase (pyrimidine biosynthesis)
relA	2822	GTP pyrophosphokinase (stringent response)
smbA	1719	uridylate kinase (pyrimidine biosynthesis)
tdk	3802	thymidine kinase
thyA	1901	thymidylate synthase A (deoxyribonucleotide biosynthesis)
thyB	2297	thymidylate synthase B (deoxyribonucleotide biosynthesis)
tmk	39	thymidylate kinase
udk	2792	uridine kinase (pyrimidine salvage)
upp	3788	uracil phosphoribosyltransferase (pyrimidine salvage)
xpt	2319	xanthine phosphoribosyltransferase (purine biosynthesis)
yaaF	23	deoxypurine kinase subunit
yaaG	24	deoxypurine kinase subunit
yabR	70	polyribonucleotide nucleotidyltransferase
yerA	713	adenine deaminase
yfkN	859	2',3'-cyclic-nucleotide 2'-phosphodiesterase
yhaM	1069	CMP-binding factor
yhcR	991	5'-nucleotidase
yirY	1144	DNA exonuclease
yjbM	1236	GTP pyrophosphokinase
yjbP	1240	diadenosine tetraphosphatase
ykkE	1377	formyltetrahydrofolate deformylase
ylbB	1565	IMP dehydrogenase
yloD	1641	guanylate kinase
ymaA	1868	ribonucleoprotein
yncB	1895	micrococcal nuclease
yncF	1899	deoxyuridine 5'-triphosphate pyrophosphatase
yosN	2165	ribonucleoside-diphosphate reductase (α subunit)
yosO	2164	ribonucleoside-diphosphate reductase (α subunit)
yosP	2161	ribonucleoside-diphosphate reductase (β subunit)
yosS	2159	deoxyuridine 5'-triphosphate nucleotidohydrolase
ypfD	2395	ribosomal protein S1 homologue
yqiB	2528	exodeoxyribonuclease VII (large subunit)
yqiC	2526	exodeoxyribonuclease VII (small subunit)
yrdF	2730	ribonuclease inhibitor
yrrU	2787	purine nucleoside phosphorylase
yumD	3302	GMP reductase
yunH	3328	allantoinase
yunL	3332	uricase
yurI	3343	ribonuclease
ywaC	3949	GTP-pyrophosphokinase

II.4 METABOLISM OF LIPIDS 77

accA	2988	acetyl-CoA carboxylase (α subunit) (long-chain fatty acid biosynthesis)
accB	2531	acetyl-CoA carboxylase (biotin carboxyl carrier subunit) (long-chain fatty acid biosynthesis)
accC	2531	acetyl-CoA carboxylase (biotin carboxylase subunit) (long-chain fatty acid biosynthesis)
acdA	3813	acyl-CoA dehydrogenase
acpA	1665	acyl carrier protein (fatty acid biosynthesis)
cdsA	1721	phosphatidate cytidylyltransferase (phospholipid biosynthesis)
dgkA	2611	diacylglycerol kinase (phospholipid biosynthesis)
fabD	1663	malonyl CoA-acyl carrier protein transacylase (fatty acid biosynthesis)
fabG	1664	3-ketoacyl-acyl carrier protein reductase (fatty acid biosynthesis)
glpQ	234	glycerophosphoryl diester phosphodiesterase (glycerol metabolism)
lcfA	2919	long chain acyl-CoA synthetase (fatty acid metabolism)
lipA	292	lipase
lipB	910	lipase
mmgA	2513	acetyl-CoA acetyltransferase
mmgB	2512	3-hydroxybutyryl-CoA dehydrogenase
mmgC	2511	acyl-CoA dehydrogenase
nap	593	carboxylesterase NA
pgsA	1762	phosphatidylglycerophosphate synthase (acidic phospholipid biosynthesis)
plsX	1662	involved in fatty acid /phospholipid synthesis
pnbA	3530	*p*-nitrobenzyl esterase
psd	249	phosphatidylserine decarboxylase (phospholipid biosynthesis)
pssA	248	phosphatidylserine synthase (phospholipid biosynthesis)
sqhC	2101	squalene-hopene cyclase (hopanoid metabolism)
ybfK	247	carboxylesterase
yclB	412	phenylacrylic acid decarboxylase
ydbM	505	butyryl-CoA dehydrogenase
ydcB	515	holo- acyl-carrier protein synthase
yfjR	871	3-hydroxyisobutyrate dehydrogenase
yhaR	1061	3-hydroxbutyryl-CoA dehydratase
yhdO	1031	1-acylglycerol-3-phosphate *O*-acyltransferase
yhdW	1038	glycerophosphodiester phosphodiesterase
yhfB	1093	3-oxoacyl- acyl-carrier protein synthase
yhfJ	1099	lipoate-protein ligase
yhfL	1100	long-chain fatty-acid-CoA ligase
yhfS	1110	acetyl-CoA *C*-acetyltransferase
yhfT	1111	long-chain fatty-acid-CoA ligase
yisP	1159	phytoene synthase
yjaX	1208	3-oxoacyl- acyl-carrier protein synthase
yjaY	1209	3-oxoacyl- acyl-carrier protein synthase
yjbW	1247	enoyl- acyl-carrier protein reductase
yjdA	1258	3-oxoacyl- acyl-carrier protein reductase
ykhA	1372	acyl-CoA hydrolase
ykwC	1465	3-hydroxyisobutyrate dehydrogenase
ymfI	1759	3-oxoacyl- acyl-carrier protein reductase
yngF	1951	3-hydroxbutyryl-CoA dehydratase
yngG	1952	hydroxymethylglutaryl-CoA lyase
yngI	1955	long-chain acyl-CoA synthetase
yngJ	1957	butyryl-CoA dehydrogenase
yocE	2089	fatty-acid desaturase
yocJ	2096	ACP phosphodiesterase
yodR	2143	butyrate-acetoacetate CoA-transferase
yodS	2144	3-oxoadipate CoA-transferase
yoxD	2019	3-oxoacyl- acyl-carrier protein reductase
yqiD	2526	geranyltranstransferase
yqiK	2514	glycerophosphodiester phosphodiesterase
yqiS	2504	phosphate butyryltransferase
yqiU	2502	branched-chain fatty-acid kinase
yqjQ	2471	ketoacyl reductase
ysiB	2917	3-hydroxbutyryl-CoA dehydratase
ytkK	3011	3-oxoacyl- acyl-carrier protein reductase
ytpA	3123	lysophospholipase
yusJ	3368	butyryl-CoA dehydrogenase
yusK	3369	acetyl-CoA *C*-acyltransferase
yusL	3372	3-hydroxyacyl-CoA dehydrogenase
yusO	3376	acyloate catabolism
yusR	3376	3-oxoacyl- acyl-carrier protein reductase
yusS	3377	3-oxoacyl- acyl-carrier protein reductase
yvaG	3450	3-oxoacyl- acyl-carrier protein reductase
yvrD	3404	ketoacyl-carrier protein reductase
ywfH	3866	3-oxoacyl- acyl-carrier protein reductase
ywhB	3853	4-oxalocrotonate tautomerase
ywiE	3822	cardiolipin synthetase
ywjE	3816	cardiolipin synthetase
ywnE	3762	cardiolipin synthase



ywpB 3743 hydroxymyristoyl-(acyl carrier protein) dehydratase
yxjD 4001 3-oxoadipate CoA-transferase
yxjE 4001 3-oxoadipate CoA-transferase

II.5 METABOLISM OF COENZYMES AND PROSTHETIC GROUPS 99

bioA 3094 adenosylmethionine-8-amino-7-oxononanoate aminotransferase (biotin biosynthesis)
bioB 3091 biotin synthetase (biotin biosynthesis)
bioD 3091 dethiobiotin synthetase (biotin biosynthesis)
bioF 3092 8-amino-7-oxononanoate synthase (biotin biosynthesis)
bioI 3089 cytochrome P450-like enzyme (biotin biosynthesis)
bioW 3094 6-carboxyhexanoate-CoA ligase (biotin biosynthesis)
dfrA 2296 dihydrofolate reductase (glycine/purine/DNA precursor synthesis, conversion of dUMP to dTMP)
dhaS 2100 aldehyde dehydrogenase
dhbA 3291 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (2,3-dihydroxybenzoate biosynthesis)
dhbB 3288 isochorismatase (2,3-dihydroxybenzoate biosynthesis)
dhbC 3291 isochorismate synthase (2,3-dihydroxybenzoate biosynthesis)
dhbE 3289 2,3-dihydroxybenzoate-AMP ligase (enterobactin synthetase component E) (2,3-dihydroxybenzoate biosynthesis)
dhbF 3287 involved in 2,3-dihydroxybenzoate biosynthesis
folA 87 dihydroneopterin aldolase (folate biosynthesis)
folC 2866 folyl-polyglutamate synthetase (folate biosynthesis)
folD 2529 methylenetetrahydrofolate dehydrogenase / methenyltetrahydrofolate cyclohydrolase (purines and amino acids biosynthesis)
folK 87 7,8-dihydro-6-hydroxymethylpterin pyrophosphokinase (dihydrofolate biosynthesis)
ggt 2004 γ-glutamyltranspeptidase (glutathione metabolism)
gsaB 943 glutamate-1-semialdehyde aminotransferase
hemA 2878 glutamyl-tRNA reductase (porphyrin biosynthesis)
hemB 2874 δ-aminolevulinic acid dehydratase (porphyrin biosynthesis)
hemC 2876 porphobilinogen deaminase (porphyrin biosynthesis)
hemD 2875 uroporphyrinogen III cosynthase (porphyrin biosynthesis)
hemE 1086 uroporphyrinogen III decarboxylase (porphyrin biosynthesis)
hemH 1087 ferrochelatase (porphyrin biosynthesis)
hemL 2873 glutamate-1-semialdehyde 2,1-aminotransferase (porphyrin biosynthesis)
hemN 2630 coproporphyrinogen III oxidase (porphyrin biosynthesis)
hemX 2877 negative effector of the concentration of HemA
hemY 1088 protoporphyrinogen IX oxidase (porphyrin biosynthesis)
menB 3149 dihydroxynapthoic acid synthetase (menaquinone biosynthesis)
menD 3151 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase / 2-oxoglutarate decarboxylase (menaquinone biosynthesis)
menE 3148 O-succinylbenzoic acid-CoA ligase (menaquinone biosynthesis)
menF 3153 menaquinone-specific isochorismate synthase (menaquinone biosynthesis)
moaB 3014 molybdopterin precursor biosynthesis
moaD 1499 molybdopterin converting factor (subunit 1)
moaE 1498 molybdopterin converting factor (subunit 2)
mobA 1495 molybdopterin-guanine dinucleotide biosynthesis
mobB 1496 molybdopterin-guanine dinucleotide biosynthesis
moeA 1497 molybdopterin biosynthesis protein
moeB 1496 molybdopterin biosynthesis protein
mtrA 2385 GTP cyclohydrolase I (tetrahydrofolate biosynthesis)
nadA 2846 quinolinate synthetase (quinolinate biosynthesis)
nadB 2349 L-aspartate oxidase (quinolinate biosynthesis)
nadC 2847 nicotinate-nucleotide pyrophosphorylase (NAD/NADP biosynthesis)
nadE 338 NH_3-dependent NAD^+ synthetase (NAD biosynthesis)
narA 3772 molybdopterin precursor biosynthesis
nasF 355 uroporphyrin-III C-methyltransferase (porphyrin biosynthesis)
nifS 2849 required for NAD biosynthesis
pabA 84 p-aminobenzoate synthase glutamine amidotransferase (subunit B) / anthranilate synthase (subunit II) (folate and tryptophan biosynthesis)
pabB 83 p-aminobenzoate synthase (subunit A) (folate biosynthesis)
pabC 65 aminodeoxychorismate lyase (folate biosynthesis)
panB 2354 ketopantoate hydroxymethyltransferase (pantothenate biosynthesis)
panC 2353 pantothenate synthetase (pantothenate biosynthesis)
panD 2352 aspartate 1-decarboxylase (pantothenate biosynthesis)
ribA 2429 GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone 4-phosphate synthase (riboflavin biosynthesis)
ribB 2429 riboflavin synthase (α subunit) (riboflavin biosynthesis)
ribC 1737 riboflavin kinase / FAD synthase (riboflavin biosynthesis)
ribG 2431 riboflavin-specific deaminase (riboflavin biosynthesis)
ribH 2428 riboflavin synthase (β subunit) (riboflavin biosynthesis)
ribT 2427 reductase (riboflavin biosynthesis)
sul 66 dihydropteroate synthase (dihydrofolate biosynthesis)
thiA 955 synthesis of the pyrimidine moiety of thiamin (thiamin biosynthesis)
thiC 3930 thiamine-phosphate pyrophosphorylase (thiamin biosynthesis)
thiD 3900 phosphomethylpyrimidine kinase (thiamin biosynthesis)
thiK 3931 hydroxyethylthiazole kinase (thiamin biosynthesis)
yaaI 26 isochorismatase
ydiA 640 thiamin-monophosphate kinase
ydiG 646 molybdopterin precursor biosynthesis
yhaV 1058 coproporphyrinogen III oxidase
yhcB 979 flavodoxin
yhfU 1112 biotin biosynthesis
yhxA [illegible]

yjbT 1245 thiamin biosynthesis
yjbU 1245 thiamin biosynthesis
yjbV 1246 phosphomethylpyrimidine kinase
ykpB 1513 thiamin biosynthesis
ykvK 1440 6-pyruvoyl tetrahydrobiopterin synthase
ykvL 1440 coenzyme PQQ synthesis
ylbO 1577 pyrimidine-thiamine biosynthesis
ylnD 1633 uroporphyrin-III C-methyltransferase
ylnF 1635 uroporphyrin-III C-methyltransferase
yloI 1642 pantothenate metabolism flavoprotein
yngH 1954 biotin carboxylase
yodC 2127 nitroreductase
yqgN 2574 5-formyltetrahydrofolate cyclo-ligase
yqjS 2469 pantothenate kinase
yrrL 2796 folate metabolism
yrrM 2795 caffeoyl-CoA O-methyltransferase
yueD 3265 sepiapterin reductase
yueJ 3261 pyrazinamidase/nicotinamidase
yueK 3260 nicotinate phosphoribosyltransferase
yuiG 3293 biotin metabolism
yurB 3335 4-hydroxybenzoyl-CoA reductase
yurC 3338 4-hydroxybenzoyl-CoA reductase
yurD 3338 4-hydroxybenzoyl-CoA reductase
yutB 3320 lipoic acid synthetase
ywaB 3950 quinone biosynthesis
ywkE 3796 protoporphyrinogen oxidase
ywoC 3755 isochorismatase

II.6 METABOLISM OF PHOSPHATE 9

phoA 1018 alkaline phosphatase A
phoB 621 alkaline phosphatase III
phoD 284 phosphodiesterase/alkaline phosphatase
phoH 2515 phosphate starvation-induced protein
xpaC 36 hydrolysis of 5-bromo-4-chloroindolyl phosphate
ybfM 248 alkaline phosphatase
ykoX 1409 alkaline phosphatase
ylaK 1549 phosphate starvation inducible protein
yngC 1947 alkaline phosphatase

II.7 METABOLISM OF SULPHUR 8

yisZ 1170 adenylylsulfate kinase
yitA 1171 sulfate adenylyltransferase
yitB 1172 phospho-adenylylsulfate sulfotransferase
ylnB 1632 sulfate adenylyltransferase
ylnC 1633 adenylylsulfate kinase
yuiH 3293 sulfite oxidase
yvgQ 3431 sulfite reductase
yvgR 3433 sulfite reductase

III INFORMATION PATHWAYS 422

III.1 DNA REPLICATION 22

dnaA 0 initiation of chromosome replication
dnaB 2965 initiation of chromosome replication / membrane attachment protein
dnaC 4158 replicative DNA helicase
dnaD 2345 initiation of chromosome replication
dnaE 2994 DNA polymerase III (α subunit)
dnaG 2603 DNA primase
dnaI 2963 primosome component (helicase loader)
dnaN 2 DNA polymerase III (β subunit)
dnaX 27 DNA polymerase III (γ and τ subunits)
holB 41 DNA polymerase III (δ' subunit)
polA 2975 DNA polymerase I
polC 1727 DNA polymerase III (α subunit)
priA 1643 primosomal replication factor Y
rnh 1677 ribonuclease H
rtp 2018 replication terminator protein
ssb 4199 single-strand DNA-binding protein
yerF 719 ATP-dependent DNA helicase
yerG 721 DNA ligase
yoqV 2192 DNA ligase
yorL 2179 DNA polymerase III (α subunit)
ypcP 2311 5'-3' exonuclease
ywpH 3740 single-strand DNA-binding protein

III.2 DNA RESTRICTION/MODIFICATION AND REPAIR 39

adaA 204 methylphosphotriester-DNA alkyltransferase / transcriptional activator of the adaAB operon
adaB 204 O^6-methylguanine-DNA methyltransferase
alkA 203 DNA-3-methyladenine glycosylase
dat 1421 O^6-methylguanine DNA alkyltransferase
dinB 606 nuclease inhibitor
dinG 2352 ATP-dependent DNA helicase
exoA 4198 3'-exo-deoxyribonuclease
mtbP 2172 modification methylase Bsu
mutL 1778 DNA mismatch repair
mutM 2972 formamidopyrimidine-DNA glycosidase
mutS 1775 DNA mismatch repair (recognition)
mutT 488 mutator protein
nth 2345 endonuclease III (DNA repair)
sms 106 DNA repair protein homologue
ung 3897 uracil-DNA glycosylase
uvrA 3612 excinuclease ABC (subunit A)
uvrB 3614 excinuclease ABC (subunit B)
uvrC 2912 excinuclease ABC (subunit C)
uvrX 2271 UV-damage repair protein
ydiO 655 DNA-methyltransferase (cytosine-specific)
ydiP 656 DNA-methyltransferase (cytosine-specific)
ydiS 660 DNA restriction
yfhQ 935 A/G-specific adenine glycosylase
yfjP 872 DNA-3-methyladenine glycosidase II
yisT 1165 nuclease inhibitor
yjcD 1255 ATP-dependent DNA helicase
yjhB 1290 mutator MutT protein
yozK 2064 DNA repair protein
yprA 2336 ATP-dependent helicase
ypvA 2329 ATP-dependent helicase
yqfS 2593 endonuclease IV
yqjH 2483 DNA-damage repair protein
yqjW 2465 ATP/GTP-binding protein
yshC 2924 DNA polymerase β
yshD 2922 DNA mismatch repair protein
ysxA 2862 DNA repair protein
yvcI 3572 mutator MutT protein
ywjD 3817 UV-endonuclease
yxlJ 3964 DNA-3-methyladenine glycosidase

III.3 DNA RECOMBINATION 17

addA 1139 ATP-dependent deoxyribonuclease (subunit A)
addB 1136 ATP-dependent deoxyribonuclease (subunit B)
recA 1764 multifunctional protein involved in homologous recombination and DNA repair (LexA-autocleavage)
recF 3 DNA repair and genetic recombination
recR 29 DNA repair and genetic recombination
ruvA 2836 Holliday junction DNA helicase
ruvB 2835 Holliday junction DNA helicase
sbcD 1143 exonuclease SbcD homologue
ylpB 1659 ATP-dependent DNA helicase
yocI 2095 ATP-dependent DNA helicase
yorK 2180 single-strand DNA-specific exonuclease
yqhH 2549 SNF2 helicase
yrrC 2808 conjugation transfer protein
yrvE 2825 single-strand DNA-specific exonuclease
ywqA 3735 SNF2 helicase

III.4 DNA PACKAGING AND SEGREGATION 10

grlA 1935 DNA gyrase-like protein (subunit A)
grlB 1933 DNA gyrase-like protein (subunit B)
gyrA 7 DNA gyrase (subunit A)
gyrB 5 DNA gyrase (subunit B)
hbs 2385 non-specific DNA-binding protein HBsu
smc 1666 chromosome segregation SMC protein homologue
smf 1682 DNA processing Smf protein homologue
topA 1683 DNA topoisomerase I
topB 476 DNA topoisomerase III
yonN 2225 HU-related DNA-binding protein

III.5 RNA SYNTHESIS 244

III.5.1 INITIATION 19

sigA 2601 RNA polymerase major sigma factor (σ^A)
sigB 522 RNA polymerase general stress sigma factor (σ^B)
sigD 1716 RNA polymerase flagella, motility, chemotaxis and autolysis sigma factor (σ^D)
sigE 1604 RNA polymerase sporulation mother cell-specific (early) sigma factor (σ^E) (SpoIIGB)
sigF 2443 RNA polymerase sporulation forespore-specific (early) sigma factor (σ^F) (SpoIIAC)
sigG 1605 RNA polymerase sporulation forespore-specific (late) sigma factor (σ^G) (SpoIIIG)
sigH 117 RNA polymerase vegetative and early stationary-phase sigma factor (σ^H) (SpoOH)
sigL 3513 RNA polymerase sigma factor (σ^L)
sigV 2769 RNA polymerase ECF-type sigma factor (σ^V)
sigW 195 RNA polymerase ECF-type sigma factor (σ^W)
sigX 2414 RNA polymerase ECF-type sigma factor (σ^X)
sigY 3970 RNA polymerase ECF-type sigma factor (σ^Y)
sigZ 2742 RNA polymerase ECF-type sigma factor (σ^Z)
spoIIIC 2701 RNA polymerase sporulation mother cell-specific (late) sigma factor (σ^K) (C-terminal half)
spoIVCB 2652 RNA polymerase sporulation mother-cell-specific (late) sigma factor (σ^K) (N-terminal half)
xpf 1324 RNA polymerase PBSX sigma factor-like
yhdM 1030 RNA polymerase ECF-type sigma factor
ykoZ 1411 RNA polymerase sigma factor
ylaC 1543 RNA polymerase ECF-type sigma factor

III.5.2 REGULATION 213

abh 1517 transcriptional regulator of transition state genes (AbrB-like)
abrB 45 transcriptional pleiotropic regulator of transition state genes (aprE, comK, ftsAZ, hpr, motAB, nprE, pbpE, rbs, spo0H, spoVG, spoOE, tycA)
acoR 883 transcriptional activator of the acetoin dehydrogenase operon (acoABCL)
ahrC 2522 transcriptional regulator of arginine metabolism expression (roc operons)
alsR 3711 transcriptional regulator of the α-acetolactate operon (alsSD)
ansR 2456 transcriptional repressor of the ansAB operon (Xre family)
araR 3485 transcriptional repressor of the arabinose operon (araABDLMNPQ)
azlB 2729 transcriptional repressor of the azlBCD operon
birA 2355 transcriptional repressor of the biotin operon (bioWAFDBI) / biotin acetyl-CoA-carboxylase synthetase
bltR 2716 transcriptional regulator of the blt operon
bmrR 2495 transcriptional activator of the bmrUR operon
ccpA 3044 transcriptional regulator involved in carbon catabolite control
cheB 1711 two-component response regulator-like [CheA] / methyl-accepting chemotaxis proteins-glutamate methylesterase
cheY 1703 two-component response regulator [CheA] involved in modulation of flagellar switch bias (chemotaxis)
citR 1020 transcriptional repressor of the citrate synthase I gene (citA)
citT 832 two-component response regulator [CitS]
codY 1690 transcriptional pleiotropic repressor (expression of srfA, comK, dpp, gabP, hut, ureABC)
comA 3253 two-component response regulator [ComP] of late competence genes / surfactin production
comK 1117 competence transcription factor (CTF), final autoregulatory control switch prior to competence development
comQ 3256 transcriptional regulator of late competence operon (comG) and surfactin expression (srfA)
ctsR 101 transcriptional repressor of class III stress genes (clpC, clpP)
degA 1163 transcriptional activator involved in the degradation of glutamine phosphoribosylpyrophosphate amidotransferase
degU 3644 two-component response regulator [DegS] involved in degradative enzyme and competence regulation (sacB, degQ, comK)
deoR 4052 transcriptional repressor of the dra/nupC/pdp operon (deoxyribonucleoside)
fnr 3831 transcriptional regulator of anaerobic genes (narK, narGHJI)
fruR 1507 transcriptional repressor of the fructose operon (fruRBA)
gerE 2904 transcriptional regulator required for expression of late spore coat genes
glcR 3739 transcriptional repressor involved in the expression of the phosphotransferase system
glcT 1456 transcriptional antiterminator essential for the expression of the ptsGHI operon
glnR 1877 transcriptional repressor of the glutamine synthetase gene (glnA)
glpP 1001 transcriptional antiterminator and control of mRNA stability of glpD
gltC 2014 transcriptional activator of the glutamate synthase operon (gltAB)
gltR 2725 transcriptional repressor of the glutamate syn...

gutR	667	transcriptional activator of the sorbitol dehydrogenase gene (*gutA*)
hpr	1073	transcriptional repressor of sporulation and extracellular proteases genes (*aprE, nprE, sin*)
hrcA	2629	transcriptional repressor of class I heat-shock genes (*dnaK, groESL*)
hutP	4040	transcriptional activator of the histidine utilization operon (*hutPHUIGM*)
iolR	4084	transcriptional repressor of the myo-inositol catabolism operon (*iolABCDEFGHIJ/iolRS*)
kdgR	2325	transcriptional repressor of the pectin utilization operon (*kdgRKAT*)
lacR	3509	transcriptional repressor of the β-galactosidase gene (*lacA*)
levR	2765	transcriptional activator of the levanase operon (*levDEFG/sacC*)
lexA	1918	transcriptional repressor of the SOS regulon
licR	3963	transcriptional regulator (antiterminator) of the lichenan operon (*licBCAH*)
licT	4012	transcriptional antiterminator required for substrate-dependent induction and catabolite repression of *bglPH*
lmrA	290	transcriptional repressor of the lincomycin operon (*lmrBA*)
lrpA	551	transcriptional Lrp-like regulator (repression of *glyA* transcription and KinB-dependent sporulation)
lrpB	552	transcriptional Lrp-like regulator (repression of *glyA* transcription and KinB-dependent sporulation)
lrpC	476	transcriptional regulator (Lrp/AsnC family)
lytR	3662	attenuator role for *lytABC* and *lytR* expression
lytT	2956	two-component response regulator [LytS] involved in the rate of autolysis
msmR	3096	transcriptional regulator (LacI family)
mta	3764	transcriptional activator of multidrug-efflux transporter genes (*bmr* and *blt*)
mtrB	2384	tryptophan operon RNA-binding attenuation protein (TRAP)
paiA	3304	transcriptional repressor of sporulation, septation and degradative enzyme genes (*aprE, nprE, phoA, sacB*)
paiB	3304	transcriptional repressor of sporulation and degradative enzyme genes
phoP	2978	two-component response regulator [PhoR] involved in phosphate regulation (*phoA, phoB, phoD, resABCDE*)
pksA	1781	transcriptional regulator of the polyketide synthase operon (*pks*)
purR	54	transcriptional repressor of the purine operon (*purEKBCLQFMNHD*)
pyrR	1618	transcriptional attenuation of the pyrimidine operon (*pyrPBCADFE*) / uracil phosphoribosyltransferase activity (minor) (pyrimidine biosynthesis)
rbsR	3700	transcriptional repressor of the ribose operon (*rbsRKDACB*)
resD	2417	two-component response regulator [ResE] involved in aerobic and anaerobic respiration (*resA, ctaA, qcrABC, fnr*)
ribR	3001	transcriptional regulator of riboflavin biosynthesis genes
rocR	4145	transcriptional activator of arginine utilization operons (*rocABC, rocDEF*)
sacT	3906	transcriptional antiterminator involved in positive regulation of *sacA* and *sacP*
sacV	532	transcriptional regulator of the levansucrase gene (*sacB*)
sacY	3942	transcriptional antiterminator involved in positive regulation of levansucrase and sucrase synthesis
senS	959	transcriptional regulator of extracellular enzyme genes (*amyE, aprE, nprE*)
sinR	2552	transcriptional regulator of post-exponential-phase responses genes (*aprE, comK, kinB, sigD, spo0A, spoIIA, spoIIE, spoIIG*)
slr	3529	transcriptional activator of competence development and sporulation genes
splA	1461	transcriptional regulator of the spore photoproduct lyase operon (*splAB*)
spo0A	2518	two-component response regulator [KinC] central for the initiation of sporulation (*spo0A, abrB, kinA, kinC, spoIIA, spoIIE, spoIIG*) (part of phosphorelay: Spo0B~P->Spo0A~P)
spo0F	3809	two-component response regulator [KinA, KinB] involved in the initiation of sporulation (part of phosphorelay: Spo0F~P->Spo0B~P)
spoIIID	3748	transcriptional regulator of σ^E- and σ^K-dependent genes
spoVT	64	transcriptional positive and negative regulator of σ^G-dependent genes
tenA	1242	transcriptional regulator of extracellular enzyme genes (*aprE, nprE, phoA, sacB*)
tenI	1243	transcriptional activator of extracellular enzyme genes
tnrA	1397	transcriptional pleiotropic regulator invoved in global nitrogen regulation (expression of *nrgAB, nasB, gabP, ureABC, glnRA*)
treR	853	transcriptional repressor of the trehalose operon (*trePAR*)
xre	1321	transcriptional repressor of PBSX genes
xylR	1891	transcriptional repressor of the xylose operon (*xylAB*)
yacF	88	transcriptional regulator (nitrogen regulation protein)
ybbB	185	transcriptional regulator (AraC/XylS family)
ybdJ	221	two-component response regulator [YbdK]
ybfI	244	transcriptional regulator (AraC/XylS family)
ybfP	251	transcriptional regulator (AraC/XylS family)
ybgA	258	transcriptional regulator (GntR family)
ycbB	267	two-component response regulator [YcbA]
ycbG	273	transcriptional regulator (GntR family)
ycbL	278	two-component response regulator [YcbM]
yccH	296	two-component response regulator [YccG]
yceK	320	transcriptional regulator (ArsR family)
ycgK	341	transcriptional regulator (LysR family)
yclA	412	transcriptional regulator (LysR family)
yclJ	426	two-component response regulator [YclK]
ycnC	438	transcriptional regulator (TetR/AcrR family)
ycnF	441	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
ycnK	449	transcriptional regulator (DeoR family)
ycsO	461	transcriptional regulator (IclR family)
ycxD	406	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
yczG	439	transcriptional regulator (ArsR family)
ydaA	467	transcriptional antiterminator (BglG family)
ydbG	499	two-component response regulator [YdbF]
ydcN	631	transcriptional regulator (phage-related) (Xre family)
ydeC	562	transcriptional regulator (AraC/XylS family)
ydeE	564	transcriptional regulator (AraC/XylS family)
ydeF	564	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
ydeL	571	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
ydeS	578	transcriptional regulator (TetR/AcrR family)
ydeT	579	transcriptional regulator (ArsR family)
ydfD	583	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
ydfI	589	two-component response regulator [YdfH]
ydgG	609	transcriptional regulator (MarR family)
ydgJ	613	transcriptional regulator (MarR family)
ydhC	616	transcriptional regulator (GntR family)
ydhQ	630	transcriptional regulator (GntR family)
yerO	732	transcriptional regulator (TetR/AcrR family)
yesN	760	two-component response regulator [YesM]
yesS	765	transcriptional regulator (AraC/XylS family)
yetL	790	transcriptional regulator (MarR family)
yezC	711	transcriptional regulator (Lrp/AsnC family)
yfiF	898	transcriptional regulator (AraC/XylS family)
yfiK	905	two-component response regulator [YfiJ]
yfiV	916	transcriptional regulator (MarR family)
yfmP	812	transcriptional regulator (MerR family)
ygaG	944	transcriptional regulator (Fur family)
yhbI	976	transcriptional regulator (MarR family)
yhcF	981	transcriptional regulator (GntR family)
yhcZ	1009	two-component response regulator [YhcY]
yhdI	1027	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
yhdQ	1033	transcriptional regulator (MerR family)
yhgD	1089	transcriptional regulator (TetR/AcrR family)
yhjM	1129	transcriptional regulator (LacI family)
yisR	1162	transcriptional regulator (AraC/XylS family)
yisV	1166	transcriptional regulator (GntR family) / aminotransferase (MocR-like)
yjdC	1270	transcriptional antiterminator (BglG family)
yjdI	1277	transcription regulation
yjmH	1308	transcriptional regulator (LacI family)
ykoG	1391	two-component response regulator [YkoH]
ykoM	1398	transcriptional regulator (MarR family)
ykuM	1485	transcriptional regulator (LysR family)
ykvE	1433	transcriptional regulator (MarR family)
ykvZ	1455	transcriptional regulator (LacI family)
ymfC	1754	transcriptional regulator (GntR family)
ynel	1923	two-component response regulator (CheY homologue)
yoaU	2045	transcriptional regulator (LysR family)
yobD	2056	transcriptional regulator (phage-related) (Xre family)
yobQ	2080	transcriptional regulator (AraC/XylS family)
yocG	2091	two-component response regulator [YocF]
yofA	2007	transcriptional regulator (LysR family)
yonR	2221	transcriptional regulator (phage-related) (Xre family)
yozA	2084	transcriptional regulator (ArsR family)
yozG	2043	transcriptional regulator
ypIP	2294	transcriptional regulator (σ-dependent)
ypoP	2287	transcriptional regulator (MarR family)
yppQ	2287	transcriptional regulator (PdB family)
ypuN	2414	negative regulator of σ^x activity
yqaE	2698	transcriptional regulator (phage-related) (Xre family)
yqcJ	2657	transcriptional regulator (ArsR family)
yqfV	2591	transcriptional regulator (Fur family)
yqhN	2543	transcriptional regulator
yqiR	2506	transcriptional regulator (σ-dependent)
yqkL	2450	transcriptional regulator (Fur family)
yraB	2755	transcriptional regulator (MerR family)
yraN	2746	transcriptional regulator (LysR family)
yrdQ	2721	transcriptional regulator (LysR family)
yrhI	2777	transcriptional regulator (TetR/AcrR family)
yrhM	2770	anti-sigma factor [σ^Y]
yrkP	2704	two-component response regulator [YrkQ]
ysiA	2918	transcriptional regulator (TetR/AcrR family)
ysmB	2904	transcriptional regulator (MarR family)
ytdP	3083	transcriptional regulator (AraC/XylS family)
ytlI	3008	transcriptional regulator (LysR family)
ytrA	3118	transcriptional regulator (GntR family)
ytsA	3113	two-component response regulator [YtsB]
ytzE	3071	transcriptional regulator (DeoR family)
yufM	3238	two-component response regulator [YufL]
yugG	3227	transcriptional regulator (Lrp/AsnC family)
yulB	3201	transcriptional regulator (DeoR family)
yurK	3345	transcriptional regulator (GntR family)
yusO	3374	transcriptional regulator (MarR family)
yusT	3377	transcriptional regulator (LysR family)
yvbA	3466	transcriptional regulator (ArsR family)
yvbU	3488	transcriptional regulator (LysR family)
yvcP	3567	two-component response regulator [YvcQ]
yvdE	3558	transcriptional regulator (LacI family)
yvdT	3540	transcriptional regulator (TetR/AcrR family)
yvfI	3509	transcriptional regulator (GntR family)
yvfU	3496	two-component response regulator [YvfT]
yvhJ	3646	transcriptional regulator
yvkB	3617	transcriptional regulator (TetR/AcrR family)
yvoA	3596	transcriptional regulator (GntR family)
yvqA	3385	two-component response regulator [YvqB]
yvqC	3394	two-component response regulator [YvqE]
yvrH	3409	two-component response regulator [YvrG]
ywaE	3945	transcriptional regulator (MarR family)
ywbI	3932	transcriptional regulator (LysR family)
ywfK	3864	transcriptional regulator (LysR family)
ywhA	3853	transcriptional regulator (MarR family)
ywoH	3748	transcriptional regulator (MarR family)
ywqM	3723	transcriptional regulator (LysR family)
ywrC	3720	transcriptional regulator (Lrp/AsnC family)
ywtF	3693	transcriptional regulator
yxaD	4109	transcriptional regulator (MarR family)
yxdJ	4072	two-component response regulator [YxdK]
yxjL	3993	two-component response regulator [YxjM]
yxjO	3991	transcriptional regulator (LysR family)
yyaG	4197	transcriptional regulator (LacI family)
yyaN	4189	transcriptional regulator (MerR family)
yybA	4183	transcriptional regulator (MarR family)
yybE	4180	transcriptional regulator (LysR family)
yycF	4154	two-component response regulator [YycG]
yydK	4122	transcriptional regulator (GntR family)

III.5.3 ELONGATION 8

greA	2791	transcription elongation factor
mfd	60	transcription-repair coupling factor
papS	2356	poly(A) polymerase
rpoA	149	RNA polymerase (α subunit)
rpoB	122	RNA polymerase (β subunit)
rpoC	126	RNA polymerase (β' subunit)
rpoE	3812	RNA polymerase (δ subunit)
yvrE	3406	RNA polymerase

III.5.4 TERMINATION 4

nusA	1732	transcription termination
nusG	118	transcription antitermination factor
rho	3804	transcriptional terminator Rho
yqhZ	2529	transcription termination

III.6 RNA MODIFICATION 19

cspR	970	rRNA methylase homolog
deaD	4016	ATP-dependent RNA helicase
miaA	1866	tRNA isopentenylpyrophosphate transferase
queA	2834	*S*-adenosylmethionine tRNA ribosyltransferase (queuosine biosynthesis)
rncS	1665	ribonuclease III
rnpA	4214	ribonuclease P (protein component)
rph	2901	ribonuclease PH
tgt	2833	tRNA-guanine transglycosylase (queuosine biosynthesis)
trmD	1675	tRNA methyltransferase
truA	153	pseudouridylate synthase I
truB	1736	tRNA pseudouridine 5S synthase
ydbR	511	ATP-dependent RNA helicase
yefA	737	RNA methyltransferase
yfjO	873	RNA methyltransferase
yfmL	816	RNA helicase
yloM	1647	RNA-binding Sun protein
yqfR	2595	ATP-dependent RNA helicase
ysgA	2931	rRNA methylase
yugI	3225	polyribonucleotide nucleotidyltransferase

III.7 PROTEIN SYNTHESIS 96

III.7.1 RIBOSOMAL PROTEINS 56

rplA	119	ribosomal protein L1 (BL1)
rplB	137	ribosomal protein L2 (BL2)
rplC	136	ribosomal protein L3 (BL3)
rplD	136	ribosomal protein L4
rplE	141	ribosomal protein L5 (BL6)
rplF	142	ribosomal protein L6 (BL8)
rplI	4163	ribosomal protein L9
rplJ	120	ribosomal protein L10 (BL5)
rplK	119	ribosomal protein L11 (BL11)
rplL	121	ribosomal protein L12 (BL9)
rplM	154	ribosomal protein L13
rplN	140	ribosomal protein L14
rplO	144	ribosomal protein L15
rplP	139	ribosomal protein L16
rplQ	150	ribosomal protein L17 (BL15)
rplR	143	ribosomal protein L18
rplS	1675	ribosomal protein L19
rplT	2952	ribosomal protein L20
rplU	2855	ribosomal protein L21 (BL20)
rplV	138	ribosomal protein L22 (BL17)
rplW	137	ribosomal protein L23
rplX	141	ribosomal protein L24 (BL23) (histone-like protein HPB12)
rpmA	2854	ribosomal protein L27 (BL24)
rpmB	1655	ribosomal protein L28
rpmC	140	ribosomal protein L29
rpmD	144	ribosomal protein L30 (BL27)
rpmE	3802	ribosomal protein L31
rpmF	1575	ribosomal protein L32
rpmG	117	ribosomal protein L33
rpmH	4215	ribosomal protein L34
rpmI	2952	ribosomal protein L35
rpmJ	148	ribosomal protein L36 (ribosomal protein B)
rpsB	1717	ribosomal protein S2
rpsC	139	ribosomal protein S3 (BS3)
rpsD	3035	ribosomal protein S4 (BS4)
rpsE	143	ribosomal protein S5
rpsF	4199	ribosomal protein S6 (BS9)
rpsG	130	ribosomal protein S7 (BS7)
rpsH	142	ribosomal protein S8 (BS8)
rpsI	154	ribosomal protein S9
rpsJ	135	ribosomal protein S10 (BS13)
rpsK	146	ribosomal protein S11 (BS11)
rpsL	130	ribosomal protein S12 (BS12)
rpsM	148	ribosomal protein S13
rpsN	142	ribosomal protein S14
rpsO	1738	ribosomal protein S15 (BS18)
rpsP	1673	ribosomal protein S16 (BS17)
rpsQ	140	ribosomal protein S17 (BS16)
rpsR	4198	ribosomal protein S18
rpsS	138	ribosomal protein S19 (BS19)
rpsT	2635	ribosomal protein S20 (BS20)
rpsU	2620	ribosomal protein S21
ybxF	129	ribosomal protein L7AE family
yhzA	965	ribosomal protein S14
ylxQ	1733	ribosomal protein L7AE family
yvyD	3631	ribosomal protein S30AE family

III.7.2 AMINOACYL-TRNA SYNTHETASES 25

alaS	2800	alanyl-tRNA synthetase
argS	3834	arginyl-tRNA synthetase
asnS	2347	asparaginyl-tRNA synthetase
aspS	2816	aspartyl-tRNA synthetase
cysS	113	cysteinyl-tRNA synthetase
gltX	111	glutamyl-tRNA synthetase
glyQ	2608	glycyl-tRNA synthetase (α subunit)
glyS	2607	glycyl-tRNA synthetase (β subunit)
hisS	2817	histidyl-tRNA synthetase
hisZ	3588	histidyl-tRNA synthetase
ileS	1613	isoleucyl-tRNA synthetase
leuS	3104	leucyl-tRNA synthetase
lysS	89	lysyl-tRNA synthetase
metS	46	methionyl-tRNA synthetase
pheS	2930	phenylalanyl-tRNA synthetase (α subunit)
pheT	2929	phenylalanyl-tRNA synthetase (β subunit)
proS	1725	prolyl-tRNA synthetase
serS	21	seryl-tRNA synthetase
thrS	2960	threonyl-tRNA synthetase (major)
thrZ	3855	threonyl-tRNA synthetase (minor)
trpS	1219	tryptophanyl-tRNA synthetase
tyrS	3037	tyrosyl-tRNA synthetase (major)
tyrZ	3946	tyrosyl-tRNA synthetase (minor)
valS	2869	valyl-tRNA synthetase
ytpR	3052	phenylalanyl-tRNA synthetase (β subunit)

III.7.3 INITIATION 6

fmt	1646	methionyl-tRNA formyltransferase
infA	148	initiation factor IF-1
infB	1733	initiation factor IF-2
infC	2952	initiation factor IF-3
rbfA	1736	ribosome-binding factor A
ykrS	1423	initiation factor eIF-2B (α subunit)

III.7.4 ELONGATION 6

efp	2538	elongation factor P
fus	131	elongation factor G

lepA 2632 GTP-binding protein
tsf 1718 elongation factor Ts
tufA 133 elongation factor Tu
ylaG 1546 GTP-binding elongation factor

III.7.5 TERMINATION 3
frr 1720 ribosome recycling factor
prfA 3797 peptide chain release factor 1
prfB 3627 peptide chain release factor 2

III.8 PROTEIN MODIFICATION 27
amhX 325 amidohydrolase
lgt 3593 prolipoprotein diacylglyceryl transferase (lipoprotein biosynthesis)
map 147 methionine aminopeptidase
pcp 287 pyrrolidone-carboxylate peptidase
ppiB 2435 peptidyl-prolyl isomerase
prkA 973 serine protein kinase
tgl 3212 transglutaminase
ybdM 224 protein kinase
ydiC 642 glycoprotein endopeptidase
ydiD 643 ribosomal-protein-alanine *N*-acetyltransferase
ydiE 643 glycoprotein endopeptidase
yfkJ 862 protein-tyrosine phosphatase
yflG 840 methionine aminopeptidase
yjcK 1261 ribosomal-protein-alanine *N*-acetyltransferase
ykrB 1526 formylmethionine deformylase
ykvY 1453 Xaa-Pro dipeptidase
yloP 1651 protein kinase
ypoP 2287 peptide methionine sulfoxide reductase
yqeT 2521 ribosomal protein L11 methyltransferase
yqhT 2539 Xaa-Pro dipeptidase
yteI 3020 protease IV
ytjP 3068 Xaa-His dipeptidase
ytvA 3105 protein kinase
ytxM 3150 prolyl aminopeptidase
yuiE 3297 leucyl aminopeptidase
ywlE 3791 protein-tyrosine-phosphatase
yxaL 4102 serine/threonine protein kinase

III.9 PROTEIN FOLDING 8
dnaK 2627 class I heat-shock protein (chaperonin)
groEL 650 class I heat-shock protein (chaperonin)
groES 650 class I heat-shock protein (chaperonin)
tig 2887 trigger factor (prolyl isomerase)
ykkC 1376 chaperonin
ykkD 1376 chaperonin
yvdR 3541 chaperonin
yvdS 3541 chaperonin

IV OTHER FUNCTIONS 289

IV.1 ADAPTATION TO ATYPICAL CONDITIONS 72
bsaA 2304 glutathione peroxidase
clpC 104 class III stress response-related ATPase (repressor of competence)
clpE 1437 ATP-dependent Clp protease-like
clpP 3545 ATP-dependent Clp protease proteolytic subunit (class III heat-shock protein)
clpQ 1688 β-type subunit of the 20S proteasome
clpX 2885 ATP-dependent Clp protease ATP-binding subunit (class III heat-shock protein)
clpY 1686 ATP-dependent Clp protease-like
csbB 930 stress response protein
cspB 984 major cold-shock protein
cspC 559 cold-shock protein
cspD 2307 cold-shock protein
cstA 2937 carbon starvation-induced protein
ctc 59 general stress protein
degQ 3256 degradative enzyme production
degR 2308 degradative enzyme production
dnaJ 2625 heat-shock protein (activation of DnaK)
dps 3135 stress- and starvation-induced gene controlled by σ^B
gbsA 3186 glycine betaine aldehyde dehydrogenase (osmoprotection)
gbsB 3184 alcohol dehydrogenase (osmoprotection)
grpE 2628 heat-shock protein (activation of DnaK)
gsiB 494 general stress protein
gspA 3944 general stress protein
hit 1076 Hit-like protein involved in cell-cycle regulation
htpG 4090 class III heat-shock protein (chaperonin)
htrA 1359 serine protease Do (heat-shock protein)
ispU 1387 activation of σ
lonA 2882 class III heat-shock ATP-dependent protease
lonB 2884 Lon-like ATP-dependent protease
mrgA 3383 metalloregulation DNA-binding stress protein
rsbR 519 positive regulator of σ^B activity (interaction with RsbS)
rsbS 520 negative regulator of σ^B activity (antagonist of RsbT)
rsbT 520 positive regulator of σ^B activity (switch protein/serine kinase [RsbS])
rsbU 521 indirect positive regulator of σ^B activity (serine phosphatase [RsbV~P])
rsbV 522 positive regulator of σ^B activity (anti-anti-sigma factor [RsbW])
rsbW 522 negative regulator of σ^B activity (switch protein/serine kinase [RsbV], anti-sigma factor [σ^B])
rsbX 523 indirect negative regulator of σ^B activity (serine phosphatase [RsbS~P])
ycdH 308 adhesion protein
ydaG 473 general stress protein
yfiQ 910 surface adhesion
ykrL 1414 heat-shock protein
ykzA 1381 general stress protein
yloA 1637 fibronectin-binding protein
yloU 1655 alkaline-shock protein
ynbA 1875 GTP-binding protein protease modulator
ynzF 1880 δ-endotoxin
yocK 2097 general stress protein
yocM 2098 small heat-shock protein
yodU 2151 capsular polysaccharide biosynthesis
yokG 2279 δ-endotoxin
ypqP 2286 capsular polysaccharide biosynthesis
ytxG 3047 general stress protein
ytxH 3047 general stress protein
ytxJ 3046 general stress protein
yveK 3529 capsular polysaccharide biosynthesis
yveL 3528 capsular polysaccharide biosynthesis
yveM 3527 capsular polysaccharide biosynthesis
yveN 3525 capsular polysaccharide biosynthesis
yveO 3524 exopolysaccharide biosynthesis
yveP 3523 capsular polysaccharide biosynthesis
yveQ 3522 capsular polysaccharide biosynthesis
yveR 3521 spore coat polysaccharide biosynthesis
yveT 3519 capsular polysaccharide biosynthesis
yvfE 3515 spore coat polysaccharide biosynthesis
yvtB 3384 serine protease Do
ywqC 3732 capsular polysaccharide biosynthesis
ywqD 3732 capsular polysaccharide biosynthesis
ywqE 3731 capsular polysaccharide biosynthesis
ywsC 3700 capsular polyglutamate biosynthesis
ywtA 3698 capsular polyglutamate biosynthesis
ywtB 3698 capsular polyglutamate biosynthesis
yyxA 4148 serine protease Do

IV.2 DETOXIFICATION 68
aadK 2736 aminoglycoside 6-adenylyltransferase
ahpC 4118 alkyl hydroperoxide reductase (small subunit)
ahpF 4119 alkyl hydroperoxide reductase (large subunit) / NADH dehydrogenase
bmrU 2493 multidrug resistance protein cotranscribed with *bmr*
cah 342 cephalosporin C deacetylase
cypA 2732 cytochrome P450-like enzyme
cypX 3603 cytochrome P450-like enzyme
katA 960 vegetative catalase 1
katB 4009 catalase 2
katX 3964 catalase
ksgA 51 dimethyladenosine transferase (kasugamycin resistance)
mmr 3857 methylenomycin A resistance protein
padC 3532 ferulate decarboxylase
penP 2048 β-lactamase
pksS 1859 hydroxylase of the polyketide produced by the *pks* cluster
sodA 2585 superoxide dismutase
sodF 2103 superoxide dismutase
tetL 4188 tetracycline resistance leader peptide
thdF 4212 thiophen and furan oxidation
tmrB 339 tunicamycin resistance
yaaN 36 toxic cation resistance
ybbE 190 β-lactamase
ybfO 250 erythromycin esterase
ybxI 229 β-lactamase
ycbJ 276 viomycin phosphotransferase
ycbR 283 toxic cation resistance protein
yceC 312 tellurium resistance protein
yceD 312 tellurium resistance protein
yceE 313 tellurium resistance protein
yceF 314 tellurium resistance protein
yceH 316 toxic anion resistance protein
ycsF 457 lactam utilization protein
ydbD 496 manganese-containing catalase
ydfB 581 antibiotic resistance protein
ydhE 618 macrolide glycosyltransferase
yerP 732 acriflavin resistance protein
yetM 790 salicylate 1-monooxygenase
yetO 792 cytochrome P450 / NADPH-cytochrome P450 reductase
yflM 836 nitric-oxide synthase
yfnC 804 fosmidomycin resistance protein
ygaF 943 thiol-specific antioxidant protein
yhjG 1122 monooxygenase
yisY 1169 chloride peroxidase
yjiB 1291 monooxygenase
yjiC 1292 macrolide glycosyltransferase
ykfA 1366 immunity to bacteriotoxins
ykkB 1375 *N*-acetyltransferase
ykoY 1410 toxic anion resistance protein
yndN 1916 fosfomycin resistance protein
yocD 2088 immunity to bacteriotoxins
yojK 2117 macrolide glycosyltransferase
yojM 2115 superoxide dismutase
yokD 2281 aminoglycoside *N3'*-acetyltransferase
yqcM 2655 arsenate reductase
yqfP 2596 penicillin tolerance
yrhJ 2776 cytochrome P450 / NADPH-cytochrome P450 reductase
yrpB 2736 2-nitropropane dioxygenase
ytgI 3017 thiol peroxidase
ytnJ 3002 nitrilotriacetate monooxygenase
yubB 3195 bacitracin resistance protein (undecaprenol kinase)
yusI 3366 arsenate reductase
yvbT 3487 alkanal monooxygenase
yvdP 3543 reticuline oxidase
ywcH 3910 monooxygenase
ywnH 3760 phosphinothricin acetyltransferase
yxeI 4062 penicillin amidase
yxeK 4061 monooxygenase
yyaR 4185 streptothricine acetyl-transferase

IV.3 ANTIBIOTIC PRODUCTION 30
pksB 1782 involved in polyketide synthesis
pksC 1783 involved in polyketide synthesis
pksD 1785 involved in polyketide synthesis
pksE 1785 involved in polyketide synthesis
pksF 1788 involved in polyketide synthesis
pksG 1789 involved in polyketide synthesis
pksH 1790 involved in polyketide synthesis
pksI 1791 involved in polyketide synthesis
pksJ 1792 involved in polyketide synthesis
pksK 1794 polyketide synthase
pksL 1808 polyketide synthase
pksM 1821 polyketide synthase
pksN 1834 polyketide synthase
pksP 1835 polyketide synthase
pksR 1850 polyketide synthase
ppsA 1997 peptide synthetase
ppsB 1990 peptide synthetase
ppsC 1982 peptide synthetase
ppsD 1974 peptide synthetase
ppsE 1963 peptide synthetase
sbo 3835 subtilosin A
sfp 408 surfactin production
srfAA 377 surfactin synthetase / competence
srfAB 387 surfactin synthetase / competence
srfAC 398 surfactin synthetase / competence
srfAD 402 surfactin synthetase / competence
sunA 2269 sublancin 168 lantibiotic antimicrobial precursor peptide
yomB 2264 bacteriocin
yukL 3282 antibiotic synthetase
yukM 3283 antibiotic synthetase

IV.4 PHAGE-RELATED FUNCTIONS 83
codV 1687 integrase/recombinase
ripX 2449 integrase/recombinase
xhlA 1346 involved in cell lysis upon induction of PBSX
xhlB 1346 hydrolysis of 5-bromo 4-chloroindolyl phosphate upon induction of PBSX (holin)
xkdC 1322 PBSX prophage
xkdD 1323 PBSX prophage
xkdE 1327 PBSX prophage
xkdF 1328 PBSX prophage
xkdG 1329 PBSX prophage
xkdH 1330 PBSX prophage
xkdI 1331 PBSX prophage
xkdJ 1331 PBSX prophage
xkdK 1332 PBSX prophage
xkdM 1333 PBSX prophage
xkdN 1334 PBSX prophage
xkdO 1334 PBSX prophage
xkdP 1338 PBSX prophage
xkdQ 1339 PBSX prophage
xkdR 1340 PBSX prophage
xkdS 1340 PBSX prophage
xkdT 1341 PBSX prophage
xkdU 1342 PBSX prophage
xkdV 1343 PBSX prophage
xkdW 1345 PBSX prophage
xkdX 1345 PBSX prophage
xkdY 1345 PBSX prophage lytic exoenzyme
xtmA 1325 PBSX terminase (small subunit)
xtmB 1325 PBSX terminase (large subunit)
xtrA 1324 PBSX prophage
ycdD 304 L-alanoyl-D-glutamate peptidase
ydcL 530 integrase
ydcM 531 immunity region protein in prophage
yhgE 1090 phage infection protein
yjbJ 1235 lytic transglycosylase
yjqB 1318 phage-related replication protein
ymaC 1863 phage-related protein
ymaH 1867 host factor-1 protein
ymfD 1755 phage-related protein
ymfE 1756 phage-related protein
yndL 1914 phage-related replication protein
yobO 2075 phage-related pre-neck appendage protein
yokA 2284 DNA recombinase
yokL 2274 phage-related protein
yolB 2272 phage-related protein
yomA 2264 holin
yomJ 2248 phage-related immunity protein
yomP 2243 phage-related protein
yomR 2242 phage-related protein
yomS 2241 phage-related lytic exoenzyme
yoqD 2200 phage-related DNA-binding protein anti-repressor
yoqZ 2190 phage-related protein
yosO 2160 phage-related endodeoxyribonuclease
yqaB 2700 phage-related protein
yqaJ 2696 phage-related protein
yqaK 2695 phage-related protein
yqaM 2694 phage-related protein
yqaO 2692 phage-related protein
yqaS 2690 phage-related terminase small subunit
yqaT 2689 phage-related terminase large subunit
yqbA 2688 phage-related protein
yqbD 2684 phage-related protein
yqbE 2683 phage-related protein
yqbH 2682 phage-related protein
yqbI 2681 phage-related protein
yqbJ 2681 phage-related protein
yqbK 2680 phage-related protein
yqbL 2679 phage-related protein
yqbM 2679 phage-related protein
yqbN 2677 phage-related protein
yqbO 2677 phage-related protein
yqbP 2672 phage-related protein
yqbQ 2671 phage-related protein
yqbR 2670 phage-related protein
yqbS 2670 phage-related protein
yqbT 2670 phage-related protein
yqcA 2669 phage-related protein
yqcC 2668 phage-related protein
yqcD 2667 phage-related protein
yqcE 2666 phage-related protein
yqxG 2666 phage-related lytic exoenzyme
yqxH 2665 holin

IV.5 TRANSPOSON AND IS 10
ydcP 533 transposon protein
ydcQ 533 transposon protein
ydcR 535 transposon protein
yddB 537 transposon protein
yddE 538 transposon protein
yddH 544 transposon protein
yefB 739 site-specific recombinase
yefC 739 resolvase
yneB 1918 resolvase
yocA 2085 transposon-related protein

IV.6 MISCELLANEOUS 26
bex 2610 GTP-binding protein
csbA 3614 putative membrane protein
csfB 36 σ^F-transcribed gene
ctaG 1564 function unknown
eag 1430 small membrane protein
ecsC 1079 function unknown
mmgE 2509 function unknown
nifZ 3027 NifS protein homologue
sapB 726 mutant activates alkaline phosphatase during sporulation independently of σ^F and σ^E
sbc 1595 small basic protein
veg 53 function unknown
yacI 102 creatine kinase
ybaL 157 ATP-binding Mrp-like protein
ycbU 287 NifS protein homologue
yerN 730 pet112-like protein
yhdP 1033 hemolysin
yhdT 1035 hemolysin
yheG 1049 calcium-binding protein
yplQ 2295 hemolysin III homologue
yqxC 2523 hemolysin-like
yrkA 2720 hemolysin-like
yrvO 2811 NifS protein homologue
yuaG 3181 epidermal surface antigen
yurV 3357 NifU protein homologue
yurW 3358 NifS protein homologue
yutI 3309 NifU protein homologue

V SIMILAR TO UNKNOWN PROTEINS 668
V.1 FROM *B. SUBTILIS* 177
V.2 FROM OTHER ORGANISMS 491
V.3 NO SIMILARITY 1053

twice. Among the duplications, we identified, as expected, the ribosomal RNA genes and their flanking regions, but also regions known to correspond to genes comprising long sequence repeats (such as *pks* and *srf*). We also found several regions that were not expected: a 182-bp repetition within the *yyaL* and *yyaO* genes; a 410-bp repetition between the *yxaK* and *yxaL* genes; an internal duplication of 174 bp inside *ydcI*; and significant duplications in the regions involved in the transcriptional control of several genes (such as 118 bp repeated three times between *yxbB* and *yxbC*). Finally, we found several repetitions at the borders of regions that might be involved in bacteriophage integration.

The most prominent duplication was a 190-bp element that was repeated 10 times in the chromosome. Multiple alignment of the ten repeats showed that they could be classified into two subfamilies with six and three copies each, plus a copy of what appears to be a chimaera. Similar sequences have also been described in the closely related species *Bacillus licheniformis*[21,22]. A striking feature of these repeats is that they are only found in half of the chromosome, at either side of the origin of replication, with five repeats on each side. Furthermore, with the exception of the most distal repeat at position 737,062, they lie in the same orientation with respect to the movement of the replication fork (Figs 2 and 3). Putative secondary structures conserved by compensatory mutations, as well as an insert in three of the copies, suggest that this element could indicate a structural RNA molecule.

Analysis at the transcription and translation level. Over 4,000 putative protein coding sequences (CDSs) have been identified, with an average size of 890 bp, covering 87% of the genome sequence (Fig. 2). We found that 78% of the genes started with ATG, 13% with TTG and 9% with GTG, which compares with 85%, 3% and 14%, respectively, in *E. coli*[8]. Fifteen genes (eight in the predicted CDSs in bacteriophage SPβ) exhibiting unusual start codons (namely ATT and CTG) were also identified through their similarities to known genes in other organisms or because they had a good GeneMark prediction (see Methods). This has not yet been substantiated experimentally. However, in the case of the gene coding for translation initiation factor 3, the similarity with its *E. coli* counterpart strongly suggests that the initiation codon is ATT, as is the case in *E. coli*.

We have not annotated CDSs that largely or entirely overlap existing genes, although such genes (for example, *comS* inside *srfAA*) certainly exist. It is also likely that some of the short CDSs present in the *B. subtilis* genome have been overlooked. For these reasons and possible sequencing errors, the estimated number of *B. subtilis* CDSs will fluctuate around the present figure of 4,100.

In several cases, in-frame termination codons or frameshifts were confirmed to be present on the chromosome (for example, an internal termination codon in *ywtF*, or the known programmed translational frameshift in *prfB*), indicating that the genes are either non-functional (pseudogenes) or subject to regulatory processes. It will therefore be of interest to determine whether these gene features are conserved in related *Bacillus* species, especially as strain 168 is derived from the Marburg strain that was subjected to X-ray irradiation[23].

A few regions do not have any identifiable feature indicating that they are transcribed: they could be 'grey holes' of the type described in *E. coli*[24]. Preliminary studies involving all regions of more than 400 bp without annotated CDSs indicated that, of ~300 such regions, only 15% were likely to be really devoid of protein-coding sequences. One of the longest such regions, located between *yfjO* and *yfjN*, is 1,628 bp long. Grey holes seem generally to be clustered near the terminus of replication. However, a grey-hole cluster located at ~600 kb might be related to the temporary chromosome partition observed during the first stages of sporulation, when a segment of about one-third of the chromosome enters the prespore, and remains the sole part of the chromosome in the prespore for a significant transition period[25].

The codon usage of *B. subtilis* CDSs was analysed using factorial correspondence analysis[17]. We found that the CDSs of *B. subtilis* could be separated into three well-defined classes (Fig. 4). Class 1 comprises the majority of the *B. subtilis* genes (3,375 CDSs), including most of the genes involved in sporulation. Class 2 (188 CDSs) includes genes that are highly expressed under exponential growth conditions, such as genes encoding the transcription and translation machineries, core intermediary metabolism, stress proteins, and one-third of genes of unknown function. Class 3 (537 CDSs) contains a very high proportion of genes of unidentified function (84%), and the members of this class have codons enriched in A + T residues. These genes are usually clustered into groups between 15 and 160 genes (for example, bacteriophage SPβ) and correspond to the A + T-rich islands described above (Fig. 1). When they are of known function, or when their products display similarity to proteins of known function, they usually correspond to functions found in, or associated with, bacteriophages or transposons, as well as functions related to the cell envelope. This includes the region *ydc/ydd/yde* (40 genes that are missing in some *B. subtilis* strains[26]), where gene products showing similarities to bacteriophage and transposon proteins are intertwined. Many of these genes are associated with virulence genes identified in pathogenic Gram-positive bacteria, suggesting that such virulence factors are transmitted horizontally among bacteria at a much higher frequency than previously thought. If we include these A + T-rich regions as possible cryptic phages, together with known bacteriophages or bacteriophage-like elements (SPβ, PBSX and the *skin* element), we find that the genome of *B. subtilis* 168 contains at least 10 such elements (Figs 2 and 3). Annotation of the corresponding regions often reveals the presence of genes that are similar to bacteriophage lytic enzymes, perhaps accounting for the observation that *B. subtilis* cultures are extremely prone to lysis.

The ribosomal RNA genes have been previously identified and

Table 1 Functional classification of the *Bacillus subtilis* protein-coding genes

The genes of known function or encoding products similar to known proteins in *B. subtilis* or in other organisms have been classified into functional categories (2,379 genes). The total number of genes in each category is indicated after the category title. Genes are listed in alphabetical order within each category, and their positions (in kilobases) on the *B. subtilis* chromosome are indicated after the gene names. A brief description is given for each gene. In some cases, interacting proteins have been indicated between brackets (for example, histidine kinases and response regulator, phosphatases and their substrates). More detailed and constantly updated information is available in the SubtiList database (see Methods). A preliminary assessment of the significance of sequence similarities was obtained through an automated procedure involving a combination between the BLAST2P probability and the percentage of amino-acid identity. Matches considered significant were re-examined manually. It should be emphasized that functions assigned to '*y*' genes are based only on sequence similarity information with the best counterparts in protein databanks. Genes whose products are only similar to other unknown proteins, or not significantly similar to any other proteins in databanks (categories V and VI), were omitted.

Figure 2 General view of the *B. subtilis* chromosome. Arrows indicate the orientation of transcription. Genes are coloured according to their classification into six broad functional categories (blue, category I; green, category II; red, category III; orange, category IV; purple, category V; pink, category VI; see Table 1). Class 2 CDSs according to codon usage analysis are indicated by oblique hatches, and class 3 CDSs are indicated by vertical hatches. Ribosomal RNA genes are coloured in yellow. Transfer RNA genes are marked by triangles. Other RNA genes are represented as white arrows. Known genes (non-'*y*' genes) are printed in bold type. Putative transcription termination sites are represented as loops. Known prophages and prophage-like elements are indicated by brown hatches on the chromosome line. The 190-bp element repeated ten times is represented by hatched boxes.

shown to be organized into ten rRNA operons, mainly clustered around the origin of replication of the chromosome (Figs 2 and 3). In addition to the 84 previously identified tRNA genes, by using the Palingol[27] and tRNAscan[28] programs, we propose four putative new tRNA loci (at 1,262 kb, 1,945 kb, 2,003 kb and 2,899 kb), specific for lysine, proline and arginine (UUU, GGG, CCU and UCU anticodons, respectively). The 10S RNA involved in degradation of proteins made from truncated mRNA has been identified (*ssrA*), as well as the RNA component of RNase P (*rnpB*) and the 4.5S RNA involved in the secretion apparatus (*scr*).

There is a strong transcription orientation bias with respect to the movement of the replication fork: 75% of the predicted genes are transcribed in the direction of replication. Plotting the density of coding nucleotides in each strand along the chromosome readily identifies the replication origin and terminus (Fig. 3). To identify putative operons, we followed ref. 29 for describing Rho-independent transcription termination sites. This yielded ~1,630 putative terminators (340 of which were bidirectional). We retained only those that were located less than 100 bp downstream of a gene, or that were considered by the program to be 'very strong' (in order to account for possible erroneous CDSs). This yielded a total of ~1,250 terminators, with a mean operon size of three genes. A similar approach to the identification of promoters is problematical, especially because at least 14 sigma factors, recognizing different promoter sequences, have been identified in *B. subtilis*. Nevertheless, the consensus of the main vegetative sigma factor (σ^A) appears to be identical to its counterpart in *E. coli* (σ^{70}): 5'-TTGACA-n_{17}-TATAAT-3'. Relaxing the constraints of the similarity to sigma-specific consensus sequences led to an extremely high number of false-positive results, suggesting that the consensus-oriented approach to the identification of promoters should be replaced by another approach[17].

Classification of gene products

Genes were classified according to ref. 14, based on the representation of cells as Turing machines in which one distinguishes between the machine and the program (Table 1). Using the BLAST2P software running against a composite protein databank compound of SWISS-PROT (release 34), TREMBL (release 3, update 1) and *B. subtilis* proteins, we assigned at least one significant counterpart with a known function to 58% of the *B. subtilis* proteins. Thus for up to 42% of the gene products, the function cannot be predicted by similarity to proteins of known function: 4% of the proteins are similar only to other unknown proteins of *B. subtilis*; 12% are similar to unknown proteins from some other organism; and 26% of the proteins are not significantly similar to any other protein in databanks. This preliminary analysis should be interpreted with caution, because only ~1,200 gene functions (30%) have been experimentally identified in *B. subtilis*. We used the '*y*' prefix in gene names to emphasize that the function has not been ascertained (2,853 '*y*' genes, representing 70%).

Regulatory systems. Transcription regulatory proteins. Helix–turn–helix proteins form a large family of regulatory proteins found in both prokaryotes and eukaryotes. There are several classes, including repressors, activators and sigma factors. Using BLAST searches, we constructed consensus matrices for helix–turn–helix proteins to analyse the *B. subtilis* protein library. We identified 18 sigma or sigma-like factors, of which nine (including a new one) are of the SigA type. We also putatively identified 20 regulators (among which 18 were products of '*y*' genes) of the GntR family, 19 regulators (15 '*y*' genes) of the LysR family, and 12 regulators (6 '*y*' genes) of the LacI family. Other transcription regulatory proteins were of the AraC family (11 members, 10 '*y*'), the Lrp family (7 members, 3 '*y*'), the DeoR family (6 members, 3 '*y*'), or additional families (such as the MarR, ArsR or TetR families). A puzzling observation is that several regulatory proteins display significant similarity to aminotransferases (seven such enzymes have been identified as showing similarity to repressors).

Two-component signal-transduction pathways. Two-component regulatory systems, consisting of a sensor protein kinase and a response regulator, are widespread among prokaryotes. We have identified 34 genes encoding response regulators in *B. subtilis*, most of which have adjacent genes encoding histidine kinases. Response regulators possess a well-conserved N-terminal phospho-acceptor domain[30], whereas their C-terminal DNA-binding domains share similarities with previously identified response regulators in *E. coli*, *Rhizobium meliloti*, *Klebsiella pneumoniae* or *Staphylococcus aureus*. Representatives of the four subfamilies recently identified in *E. coli*

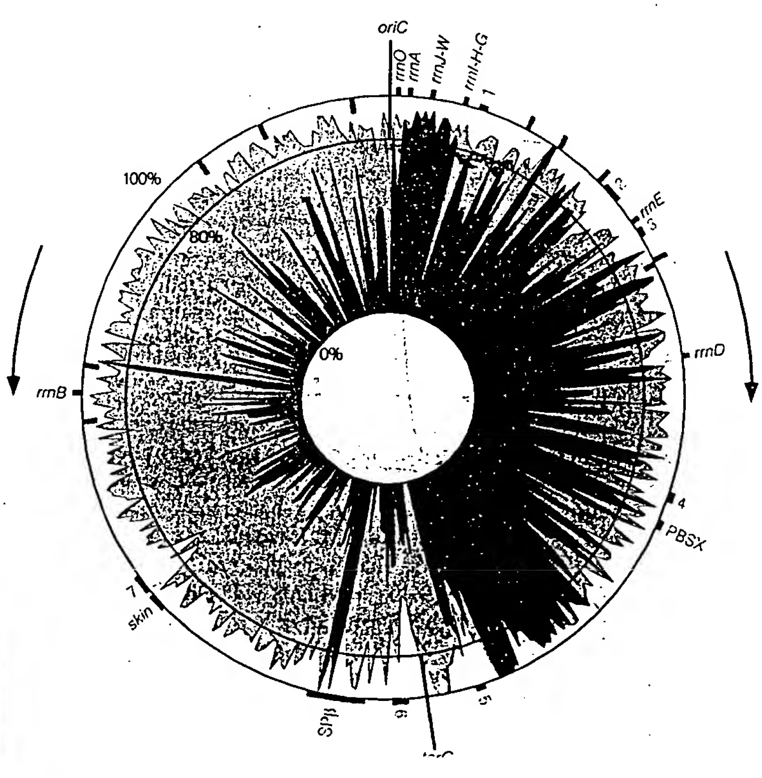


Figure 3 Density of coding nucleotides along the *B. subtilis* chromosome. Yellow stands for the density of coding nucleotides in both strands of the sequence; red indicates the density of coding nucleotides in the clockwise strand (nucleotides involved in genes transcribed in the clockwise orientation). The movement of the replication forks is represented by arrows. Ribosomal RNA operons are indicated by brown boxes. Known prophages and prophage-like elements are represented as blue lines. The 190-bp element repeated ten times is represented by green lines.

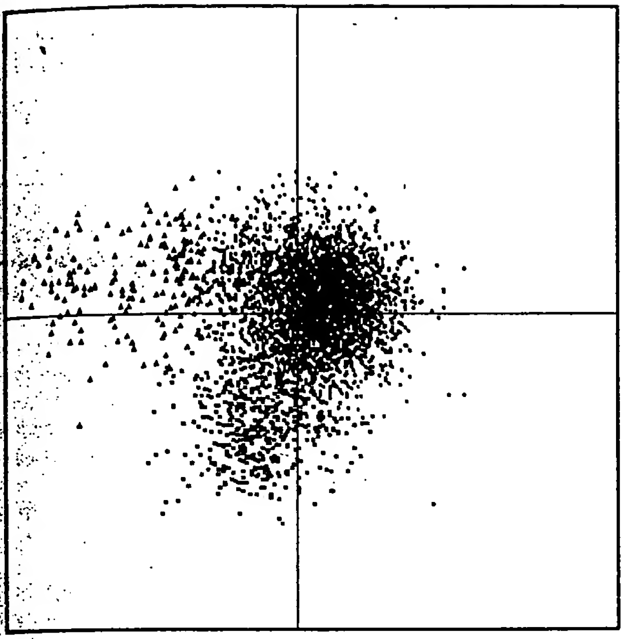


Figure 4 Factorial correspondence analysis of codon usage in the *B. subtilis* CDSs. Red dots, genes from class 1; green triangles, genes from class 2; blue crosses, genes from class 3. Class 2 contains genes coding for the translation and transcription machineries, and genes of the core intermediary metabolism. Class 3 genes correspond to codons strongly enriched in A or T in the wobble position; they generally belong to prophage-like inserts in the genome.

(OmpR, FixJ, CitB and LytR) have been identified in *B. subtilis*. In a fifth subfamily, CheY, the DNA-binding domain is absent. The DNA-binding domain of a single *B. subtilis* response regulator, YesN, shares similarity with regulatory proteins of the AraC family.

Quorum sensing. The *B. subtilis* genome contains 11 aspartate phosphatase genes, whose products are involved in dephosphorylation of response regulators, that do not seem to have counterparts in Gram-negative bacteria such as *E. coli*. Downstream from the corresponding genes are some small genes, called *phr*, encoding regulatory peptides that may serve as quorum sensors[32]. Seven *phr* genes have been identified so far, including three new genes (*phrG*, *phrI* and *phrK*).

Protein secretion. It is known that *B. subtilis* and related *Bacillus* species, in particular *B. licheniformis* and *B. amyloliquefaciens*, have a high capacity to secrete proteins into the culture medium. Several genes encoding proteins of the major secretion pathway have been identified: *secA*, *secD*, *secE*, *secF*, *secY*, *ffh* and *ftsY*. Surprisingly, there is no gene for the SecB chaperone. It is thought that other chaperone(s) and targeting factor(s), such as Ffh and FtsY, may take over the SecB function. Further, although there is only one such gene in *E. coli*, five type I signal peptidase genes (*sipS*, *sipT*, *sipU*, *sipV* and *sipW*) have been found[33]. The *lsp* gene, encoding a type II signal peptidase required for processing of lipo-modified precursors, was also identified. PrsA, located at the outer side of the membrane, is important for the refolding of several mature proteins after their translocation through the membrane.

Other families of proteins. ABC transporters were the most frequent class of proteins found in *B. subtilis*. They must be extremely important in Gram-positive bacteria, because they have an envelope comprising a single membrane. ABC transporters will therefore allow such bacteria to escape the toxic action of many compounds. We propose that 77 such transporters are encoded in the genome. In general they involve the interaction of at least three gene products, specified by genes organized into an operon. Other families comprised 47 transport proteins similar to facilitators (and perhaps sometimes part of the ABC transport systems), 18 amino-acid permeases (probably antiporters), and at least 16 sugar transporters belonging to the PEP-dependent phosphotransferase system.

General stress proteins are important for the survival of bacteria under a variety of environmental conditions. We identified 43 temperature-shock and general stress proteins displaying strong similarity to *E. coli* counterparts.

Missing genes. Histone-like proteins such as HU and H-NS have been identified in *E. coli*. We found that *B. subtilis* encodes two putative histone-like proteins that show similarity to *E. coli* HU, namely HBsu and YonN, but found no homologue to H-NS. It is known that the *hbs* gene encoding HBsu is essential, but we do not expect the *yonN* gene to be essential because it is present in the SPβ prophage. IHF is similar to HU, and it is not known whether HBsu plays a similar role to that of IHF in *E. coli*. Similarly, no protein similar to FIS could be found.

Genes encoding products that interact with methylated DNA, such as *seqA* in *E. coli*, involved in the regulation of replication initiation timing, or *mutH*, the endonuclease recognizing the newly synthesized strand during mismatch repair at hemi-methylated

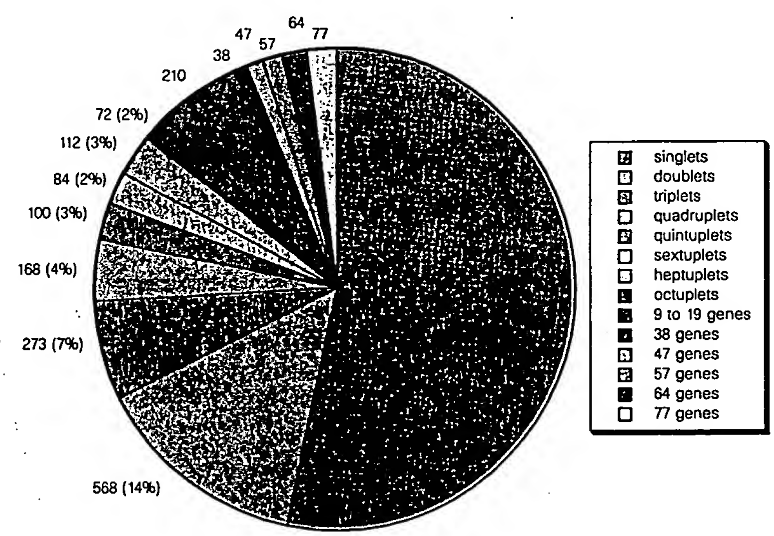


Figure 5 Gene paralogue distribution in the genome of *B. subtilis*. Each *B. subtilis* protein has been compared with all other proteins in the genome, using a Smith–Waterman algorithm. The baseline is established by making a similar comparison using 100 independent random shuffles of the protein sequence (Z-score > 13).

GATC sites, are also missing. This is in line with the absence of known methylation in *B. subtilis*, equivalent to Dam methylation in *E. coli*. Similarly, *E. coli sfiA*, encoding an inhibitor of FtsZ action in the SOS response, has no counterpart in *B. subtilis*. In contrast, *B. subtilis* replication initiation-specific genes, such as *dnaB* and *dnaD*, are missing in *E. coli*. The exact counterpart of the *E. coli mukB* gene, involved in chromosome partitioning, does not exist in *B. subtilis*, but genes *spo0J* and *smc* (Smc is weakly similar to MukB), which are suggested to be involved in partitioning of the *B. subtilis* chromosome, are missing in *E. coli*.

Turnover of mRNA is controlled in *E. coli* by a 'degradosome' comprising RNase E. It has a counterpart in *B. subtilis*, but we failed to find a clear homologue of RNase E in this organism. Whether this is related to the role of ribosomal protein S1 as an RNA helicase involved in mRNA turnover in *E. coli* requires further investigation. In particular, a homologue of *rpsA* (S1 structural gene), *ypfD*, might be involved in a structure homologous to the degradosome[34].

Structurally unrelated genes of similar function. Several genes encode products that have similar functions in *E. coli* and *B. subtilis*, but have no evident common structure. This is the case for the helicase loader genes, *E. coli dnaC* and *B. subtilis dnaI*; the genes coding for the replication termination protein, *E. coli tus* and *B. subtilis rtp*; and the division topology specifier genes, *E. coli minE* and *B. subtilis divIVA*. The situation may even be more complex in multisubunit enzymes: *B. subtilis* synthesizes two DNA polymerase III α chains, one having 3′–5′ proofreading exonuclease activity (PolC) and the other without the exonuclease activity (DnaE); in *E. coli*, only the latter exists. *E. coli* DNA polymerase II is structurally related to DNA polymerase α of eukaryotes, whereas *B. subtilis* YshC is related to DNA polymerase β.

Metabolism of small molecules

The type and range of metabolism used for the interconversion of low-molecular-weight compounds provide important clues to an organism's natural environment(s) and its biological activity. Here we briefly outline the main metabolic pathways of *B. subtilis* before the reconstruction of these pathways *in silico*, the correlation of genes with specific steps in the pathway, and ultimately the prediction of patterns of gene expression.

Intermediary metabolism. It has long been known that *B. subtilis* can use a variety of carbohydrates. As expected, it encodes an Embden–Meyerhof–Parnas glycolytic pathway, coupled to a functional tricarboxylic acid cycle. Further, *B. subtilis* is also able to grow anaerobically in the presence of nitrate as an electron acceptor. This metabolism is, at least in part, regulated by the FNR protein, binding to sites upstream of at least eight genes (four sites experimentally confirmed and four putative sites). A noteworthy feature of *B. subtilis* metabolism is an apparent requirement of branched short-chain carboxylic acids for lipid biosynthesis[35]. Branched-chain 2-keto acid decarboxylase activity exists and may be linked to a variety of genes, suggesting that *B. subtilis* can synthesize and utilize linear branched short-chain carboxylic acids and alcohols.

Amino-acid and nucleotide metabolism. Pyrimidine metabolism of *B. subtilis* seems to be regulated in a way fundamentally different from that of *E. coli*, as it has two carbamylphosphate synthetases (one specific for arginine synthesis, the other for pyrimidine). Additionally, the aspartate transcarbamylase of *B. subtilis* does not act as an allosteric regulator as it does in *E. coli*. As in other microorganisms, pyrimidine deoxyribonucleotides are synthesized from ribonucleoside diphosphates, not triphosphates. The cytidine diphosphate required for DNA synthesis is derived from either the salvage pathway of mRNA turnover or from the synthesis of phospholipids and components of the cell wall. This means that polynucleotide phosphorylase is of fundamental importance in nucleic acid metabolism, and may account for its important role in competence[36]. Two ribonucleoside reductases, both of class I, NrdEF type, are encoded by the *B. subtilis* chromosome, in one case from within the SPβ genome. In this latter case, the gene corresponding to the large subunit both contains an intron and codes an intein (V.L., unpublished data). The gene of the small subunit of this enzyme also contains an intron, encoding an endonuclease, as was found for the homologue in bacteriophage T4.

By similarity with genes from other organisms, there appears to be, in addition to genes involved in amino-acid degradation (such as the *roc* operon, which degrades arginine and related amino acids), a large number of genes involved in the degradation of molecules such as opines and related molecules, derived from plants. This is also in line with the fact that *B. subtilis* degrades polygalacturonate and suggests that, in its biotope, it forms specific relations with plants.

Secondary metabolism. In addition to many genes coding for degradative enzymes, almost 4% of the *B. subtilis* genome codes for large multifunctional enzymes (for example, the *srf*, *pps* and *pks* loci), similar to those involved in the synthesis of antibiotics in other genera of Gram-positive bacteria such as *Streptomyces*. Natural isolates of *B. subtilis* produce compounds with antibiotic activity, such as surfactin, fengycin and difficidin, that can be related to the above-mentioned loci. This bacterium therefore provides a simple and genetically amenable model in which to study the synthesis of antibiotics and its regulation. These pathways are often organized in very long operons (for example, the *pks* region spans 78.5 kb, about 2% of the genome). The corresponding sequences are mostly located near the terminus of replication, together with prophages and prophage-like sequences.

Paralogues and orthologues

It is important to relate intermediary metabolism to genome structure, function and evolution. We therefore compared the *B. subtilis* proteins with themselves, as well as with proteins from known complete genomes, using a consistent statistical method that allows the evaluation of unbiased probabilities of similarities between proteins[37,38]. For Z-scores higher than 13, the number of proteins similar to each given protein does not vary, indicating that this cut-off value identifies sets of proteins that are significantly similar.

Families of paralogues. Many of the paralogues constitute large families of functionally related proteins, involved in the transport of compounds into and out of the cell, or involved in transcription regulation. Another part of the genome consists of gene doublets (568 genes), triplets (273 genes), quadruplets (168 genes) and quintuplets (100 genes). Finally, about half of the genome is made of genes coding for proteins with no apparent paralogues (Fig. 5). No large family comprises only proteins without any similarity to proteins of known function.

The process by which paralogues are generated is not well understood, but we might find clues by studying some of the duplications in the genome. Several approximate DNA repetitions, associated with very high levels of protein identity, were found mainly within regions putatively or previously identified as prophages. This is in line with previous observations about PBSX and the *skin* element[39,40], and suggests that these prophage-like elements share a common ancestor and have diverged relatively recently. In addition, several protein duplications are in genes that are located very close to each other, such as *yukL* and *dhbF* (the corresponding proteins are 65% identical in an overlap of 580 amino acids), *yugJ* and *yugK* (proteins 73% identical), *yxjG* and *yxjH* (proteins 70% identical), and the entire *opuB* operon, which is duplicated 3 kb away (*opuC* operon, yielding ~80% of amino-acid identity in the corresponding proteins).

The study of paralogues showed that, as in other genomes, a few classes of genes have been highly expanded. This argues against the idea of the genome evolving through a series of duplications of ancestral genomes, but rather for the idea of genes as living organisms, subject to evolutionary constraints, some being sub-

mitted to expansion and natural selection, and others to local duplications of DNA regions.

Among paralogue doublets, some were unexpected, such as the three aminoacyl tRNA synthetases doublets (*hisS* (2,817 kb) and *hisZ* (3,588 kb); *thrS* (2,960 kb) and *thrZ* (3,855 kb); *tyrS* (3,036 kb) and *tyrZ* (3,945 kb)) or the two *mutS* paralogues (*mutS* and *yshD*). This latter situation is similar to that found in *Synechocystis*. In the case of *B. subtilis*, the presence of two MutS proteins could indicate that there are two different pathways for long-patch mismatch repair, possibly a consequence of the active genetic transformation mechanism of *B. subtilis*.

Families of orthologues. Because *Mycoplasma* spp. are thought to be derived from Gram-positive bacteria similar to *B. subtilis*, we compared the *B. subtilis* genome with that of *M. genitalium*. Among the 450 genes encoded by *M. genitalium*, the products of 300 are similar to proteins of *B. subtilis*. Among the 146 remaining gene products, a further 3 are similar to proteins of other *Bacillus* species, and 9 to proteins of other Gram-positive bacteria; 25 are similar to proteins of Gram-negative bacteria; and 19 are similar to proteins of other *Mycoplasma* spp. This leaves only 90 genes that would be specific to *M. genitalium* and might be involved in the interaction of this organism with its host.

The *B. subtilis* genome is similar in size to that of *E. coli*. Because these bacteria probably diverged more than one billion years ago, it is of evolutionary value to investigate their relative similarity. About 1,000 *B. subtilis* genes have clear orthologous counterparts in *E. coli* (one-quarter of the genome). These genes did not belong either to the prophage-like regions or to regions coding for secondary metabolism (~15% of the *B. subtilis* genome). This indicates that a large fraction of these genomes shared similar functions. At first sight, however, it seems that little of the operon structure has been conserved. We nevertheless found that ~100 putative operons or parts of operons were conserved between *E. coli* and *B. subtilis*. Among these, ~12 exhibited a reshuffled gene order (typically, the arabinose operon is *araABD* in *B. subtilis* and *araBAD* in *E. coli*). In addition to the core of the translation and transcription machinery, we identified other classes of operons that were well conserved between the two organisms, including major integrated functions such as ATP synthesis (*atp* operon) and electron transfer (*cta* and *qox* operons). As well as being well preserved, the murein biosynthetic region was partly duplicated, allowing creation of part of the genes required for the sporulation division machinery[41]. The amino-acid biosynthesis genes differ more in their organization: the *E. coli* genes for arginine biosynthesis are spread throughout the chromosome, whereas the arginine biosynthesis genes of *B. subtilis* form an operon. The same is true for purine biosynthetic genes. Genes responsible for the biosynthesis of coenzymes and prosthetic groups in *B. subtilis* are often clustered in operons that differ from those found in *E. coli*. Finally, several operons conserved in *E. coli* and *B. subtilis* correspond to unknown functions, and should therefore be priority targets for the functional analysis of these model genomes.

Comparison with *Synechocystis* PCC6803 revealed about 800 orthologues. However, in this case the putative operon structure is extremely poorly conserved, apart from four of the ribosomal protein operons, the *groES–groEL* operon, *yfnHG* (respectively in *Synechocystis rfbFG*), *rpsB-tsf*, *ylxS-nusA-infB*, *asd-dapGA-ymfA*, [illegible]*AB*, *efp-accB*, *grpE-dnaK*, *yurXW*. The nine-gene *atp* operon of *B. subtilis* is split into two parts in *Synechocystis*: *atpBE* and *atpIHGFDAC*.

Conclusion

The biochemistry, physiology and molecular biology of *B. subtilis* have been extensively studied over the past 40 years. In particular, *B. subtilis* has been used to study postexponential phase phenomena such as sporulation and competence for DNA uptake. The genome sequences of *E. coli* and *B. subtilis* provide a means of studying the evolutionary divergence, one billion years ago, of eubacteria into the Gram-positive and Gram-negative groups. The availability of powerful genetic tools will allow the *B. subtilis* genome sequence data to be exploited fully within the framework of a systematic functional analysis program, undertaken by a consortium of 19 European and 7 Japanese laboratories coordinated by S. D. Ehrlich (INRA, Jouy-en-Josas, France) and by N. Ogasawara and H. Yoshikawa (Nara Institute of Science and Technology, Nara, Japan). □

Methods

Genome cloning and sequencing. An international consortium was established to sequence the genome of *B. subtilis* strain 168 (refs 9, 10, 42). At its peak, 25 European, seven Japanese and one Korean laboratory participated in the program, together with two biotechnology companies. Five contiguous DNA regions totalling 0.94 Mb, and two additional regions of 0.28 and 0.14 Mb, were sequenced by the Japanese partners, while the European partners sequenced a total of 2.68 Mb. A few sequences from strain 168 published previously were not resequenced when long overlaps did not indicate differences.

A major technical difficulty was the inability to construct in *E. coli* gene banks representative of the entire *B. subtilis* chromosome using vectors that have proved efficient for other sources of bacterial DNA (such as bacteriophage or cosmid vectors). This was due to the generally very high level of expression of *B. subtilis* genes in *E. coli*, leading to toxic effects. This limitation was overcome by: cloning into a variety of vectors[9,43,44]; using an *E. coli* strain maintaining low-copy number plasmids[44]; using an integrative plasmid/marker rescue genome-walking strategy[44]; and *in vitro* amplification using polymerase chain reaction (PCR) techniques[45,46].

Although cloning vectors were used in the early stages as templates for sequencing reactions, they were largely superseded in the later stages by long-range and inverse PCR techniques. To reduce sequencing errors resulting from PCR amplification artefacts, at least eight amplification reactions were performed independently and subsequently pooled. The various sequencing groups were free to choose their own strategy, except that all DNA sequences had to be determined entirely on both strands.

Sequence annotation and verification. The sequences were annotated by the groups, and sent to a central depository at the Institut Pasteur[14]. The Japanese sequences were also sent there through the Japanese depository at the Nara Institute of Science and Technology. The same procedures were used to identify CDSs and to detect frameshifts. They were embedded within a cooperative computer environment dedicated to automatic sequence annotation and analysis[39]. In a first step, we identified in all six possible frames the open reading frames (ORFs) that were at least 100 codons in length. In a second step, three independent methods were used: the first method used the GeneMark coding-sequence prediction method[47] together with the search for CDSs preceded by typical translation initiation signals (5′-AAGGAGGTG-3′), located 4–13 bases upstream of the putative start codons (ATG, TTG or GTG); the second method used the results of a BLAST2X analysis performed on the entire *B. subtilis* genome against the non-redundant protein databank at the NCBI; and the third method was based on the distribution of non-overlapping trinucleotides or hexanucleotides in the three frames of an ORF[48].

In general, frameshifts and missense mutations generating termination codons or eliminating start codons are relatively easy to detect. We shall devise a procedure for detecting another type of error, GC instead of CG or vice versa, which are much more difficult to identify. It should be noted that putative frameshift errors should not be corrected automatically. The sequences of the flanking regions of a 500-bp fragment centred around a putative error were sent to an independent verification group, which performed PCR amplifications using chromosomal DNA as template, and sequenced the corresponding DNA products.

Organization and accessibility of data. The *B. subtilis* sequence data have been combined with data from other sources (biochemical, physiological and genetic) in a specialized database, SubtiList[49], available as a Macintosh or Windows stand-alone application (4th Dimension runtime) by anonymous FTP at ftp://ftp.pasteur.fr/pub/GenomeDB/SubtiList. SubtiList is also accessible through a World-Wide Web server at http://www.pasteur.fr/Bio/SubtiList.html,

where it has been implemented on a UNIX system using the Sybase relational database management system. A completely rewritten version of SubtiList is in preparation to facilitate browsing of the information of the whole chromosome. Flat files of the whole DNA and protein sequences in EMBL and FASTA format will be made available at the above ftp address. Another *B. subtilis* genome database is also under development at the Human Genome Center of Tokyo University (http://www.genome.ad.jp), and SubtiList will also be available there.

Received 16 July; 29 September 1997.

1. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512 (1995).
2. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403 (1995).
3. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3, 109–136 (1996).
4. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073 (1996).
5. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24, 4420–4449 (1996).
6. Goffeau, A. *et al.* The yeast genome directory. *Nature* 387, 5–105 (1997).
7. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388, 539–547 (1997).
8. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1462 (1997).
9. Kunst, F., Vassarotti, A. & Danchin, A. Organization of the European *Bacillus subtilis* genome sequencing project. *Microbiology* 389, 84–87 (1995).
10. Ogasawara, N. & Yoshikawa, H. The systematic sequencing of the *Bacillus subtilis* genome in Japan. *Microbiology* 142, 2993–2994 (1996).
11. Harwood, C. R. *Bacillus subtilis* and its relatives: molecular biological and industrial workhorses. *Trends Biotechnol.* 10, 247–256 (1992).
12. Stragier, P. & Losick, R. Molecular genetics of sporulation in *Bacillus subtilis*. *Annu. Rev. Genet.* 30, 297–341 (1996).
13. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. *Trends Genet.* 12, 150–155 (1996).
14. Moszer, I., Kunst, F. & Danchin, A. The European *Bacillus subtilis* genome sequencing project: current status and accessibility of the data from a new World Wide Web site. *Microbiology* 142, 2987–2991 (1996).
15. Franks, A. H., Griffiths, A. A. & Wake, R. G. Identification and characterization of new DNA replication terminators in *Bacillus subtilis*. *Mol. Microbiol.* 17, 13–23 (1995).
16. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13, 660–665 (1996).
17. Hénaut, A. & Danchin, A. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (eds Neidhardt, F. *et al.*) 2047–2066 (ASM, Washington DC, 1996).
18. Nussinov, R. The universal dinucleotide asymmetry rules in DNA and amino acid codon choice. *Nucleic Acids Res.* 17, 237–244 (1981).
19. Karlin, S., Burge, C. & Campbell, A. M. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* 20, 1363–1370 (1992).
20. Burge, C., Campbell, A. M. & Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA* 89, 1358–1362 (1992).
21. Kasahara, Y., Nakai, S. & Ogasawara, H. Sequence analysis of the 36-kb region between *gntZ* and *trnY* genes of *Bacillus subtilis* genome. *DNA Res.* 4, 155–159 (1997).
22. Presecan, E. *et al.* The *Bacillus subtilis* genome from *gerBC* (311°) to *licR* (334°). *Microbiology* 143, 3313–3328 (1997).
23. Burkholder, P. R. & Giles, N. H. Induced biochemical mutations in *Bacillus subtilis*. *Am. J. Bot.* 33, 345–348 (1947).
24. Daniels, D. L., Plunkett, G. III, Burland, V. & Blattner, F. R. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* 257, 771–778 (1992).
25. Wu, L. J. & Errington, J. *Bacillus subtilis* SpoIIIE protein required for DNA segregation during asymmetric cell division. *Science* 264, 572–575 (1994).
26. Itaya, M. Stability and asymmetric replication of the *Bacillus subtilis* 168 chromosome structure. *J. Bacteriol.* 175, 741–749 (1993).
27. Billoud, B., Kontic, M. & Viari, A. Palingol: a declarative programming language to describe acids' secondary structures and to scan sequence database. *Nucleic Acids Res.* 24, 1395–1403
28. Fichant, G. A. & Burks, C. Identifying potential tRNA genes in genomic DNA sequences. *J.* 220, 659–671 (1991).
29. d'Aubenton Carafa, Y., Brody, E. & Thermes, C. Prediction of rho-independent *Escheri* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Bi* 835–858 (1990).
30. Stock, J. B., Surette, M. G., Levitt, M. & Park, P. in *Two-Component Signal Transduction* (eds Hoc & Silhavy, T. J.) 25–51 (ASM, Washington DC, 1995).
31. Mizuno, T. Compilation of all genes encoding two-component phosphotransfer signal transdu the genome of *Escherichia coli*. *DNA Res.* 4, 161–168 (1997).
32. Perego, M., Glaser, P. & Hoch, J. A. Aspartyl-phosphate phosphatases deactivate the res regulator components of the sporulation signal transduction system in *Bacillus subtilis*. *Microbiol.* 19, 1151–1157 (1996).
33. Tjalsma, H. *et al.* *Bacillus subtilis* contains four closely related type I signal peptidases with overl substrate specificities: constitutive and temporally controlled expression of different *sip* genes. *Chem.* 272, 25983–25992 (1997).
34. Danchin, A. Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res.* 4, 9–18 (1997).
35. Suutari, M. & Laakso, S. Unsaturated and branched chain-fatty acids in temperature adaptation *Bacillus subtilis* and *Bacillus megaterium*. *Biochim. Biophys. Acta* 1126, 119–124 (1992).
36. Luttinger, A., Hahn, J. & Dubnau, D. Polynucleotide phosphorylase is necessary for compe development in *Bacillus subtilis*. *Mol. Microbiol.* 19, 343–356 (1996).
37. Landès, C., Hénaut, A. & Risler, J.-L. A comparison of several similarity indices used in classification of protein sequences: a multivariate analysis. *Nucleic Acids Res.* 20, 3631–3637 (199
38. Glémet, E. & Codani, J.-J. LASSAP, a LArge Scale Sequence compArison Package. *Comput. A Biosci.* 13, 137–143 (1997).
39. Médigue, C., Moszer, I., Viari, A. & Danchin, A. Analysis of a *Bacillus subtilis* genome fragment usin co-operative computer system prototype. *Gene* 165, GC37–GC51 (1995).
40. Krogh, S., O'Reilly, M., Nolan, N. & Devine, K. M. The phage-like element PBSX and part of the element, which are resident at different locations on the *Bacillus subtilis* chromosome, are hi homologous. *Microbiology* 142, 2031–2040 (1996).
41. Daniel, R. A., Drake, S., Buchanan, C. E., Scholle, R. & Errington, J. The *Bacillus subtilis spoVD* ge encodes a mother-cell-specific penicillin-binding protein required for spore morphogenesis. *J. Mo Biol.* 235, 209–220 (1994).
42. Anagnostopoulos, C. & Spizizen, J. Requirements for transformation in *Bacillus subtilis*. *J. Bacter* 81, 741–746 (1961).
43. Azevedo, V. *et al.* An ordered collection of *Bacillus subtilis* DNA segments cloned in yeast artifici chromosomes. *Proc. Natl Acad. Sci. USA* 90, 6047–6051 (1993).
44. Glaser, P. *et al.* *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325 to 333°. *Mol. Microbiol.* 10, 371–384 (1993).
45. Ogasawara, N., Nakai, S. & Yoshikawa, H. Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Res.* 1, 1–14 (1994).
46. Sorokin, A. *et al.* A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. *Genome Res.* 6, 448–453 (1996).
47. Borodovsky, M. & McIninch, J. GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.* 17, 123–133 (1993).
48. Fichant, G. A. & Quentin, Y. A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res.* 23, 2900–2908 (1995).
49. Moszer, I., Glaser, P. & Danchin, A. SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology* 141, 261–268 (1995).

Acknowledgements. We thank C. Anagnostopoulos, R. Dedonder and J. Hoch for their pioneering efforts, and A. Bairoch for advice in annotating *B. subtilis* protein data. The main funding of the European network was provided by the European Commission under the Biotechnology program. The Japanese project was included in the Human Genome Program, and supported by a research grant from the Ministry of Education, Science and Culture, and the Proposal-Based Advanced Industrial Technology R&D Program from New Energy and Industrial Technology Development Organization. The Swiss and Korean projects were funded by the Swiss National Fund and the Korean government, respectively. An industrial platform was set up to facilitate contacts between participants of the European consortium and some European biotechnology companies: DuPont de Nemours (France, USA), Frimond (Belgium), Genencor (Finland, USA), Gist Brocades (The Netherlands), Glaxo-Wellcome (UK, Italy), Hoechst Marion Roussel (France, Germany), F. Hoffmann-La Roche AG (Switzerland), Novo Nordisk (Denmark), SmithKline Beecham (UK).

Correspondence and requests for materials should be addressed to F.K. (e-mail: fkunst@pasteur.fr), N.O. (nogasawa@bs.aist-nara.ac.jp), H.Y. (hyoshika:bs.aist-nara.ac.jp) or A.D. (adanchin@pasteur.fr). The sequence has been deposited in EMBL/GenBank/DDBJ with accession numbers from Z99104 to Z99124.

OU Streptococcus Pyogenes Sequence Blast Server Results

```
TBLASTN 1.3.9 [29-Oct-93]
Reference:  Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers,
and David J. Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol.
215:403-410.
Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.
Query=  mrwypwlrpd feklvasyqa grghhalliq alpgmgddal iyalsryllc qqpqghkscg
        (274 letters)
Database:  /strep/abi/spphrap/auto_strep
           139 sequences; 1,816,476 total letters.
Searching..................................................done
                                                                 Smallest
                                                                  Poisson
                                                      Reading  High  Probability
Sequences producing High-scoring Segment Pairs:         Frame Score  P(N)      N
Contig218                                                 +3   122  3.3e-10   1
Contig203                                                 +1   100  4.0e-07   1
Contig215                                                 +3    49  0.95      2
Contig173                                                 -1    42  0.99      4
>Contig218
          Length = 36,214
  Plus Strand HSPs:
 Score = 122 (56.3 bits), Expect = 3.3e-10, P = 3.3e-10
 Identities = 31/97 (31%), Positives = 47/97 (48%), Frame = +3
Query:     2 CRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLT 61
             C  C + +   T+      +        +  GVD +R++ +K           KV  + +  +L+
Sbjct: 33567 CNQCDICRDITNGSLEDVIEIDAASNNGVDEIRDIRDKSTYAPSRATYKVYIIDEVHMLS 33746
Query:    62 DAAANALLKTLEEPPAETWFFLATREPERLLATLRSR 98
               A NALLKTLEEP      F LAT E  ++ AT+ SR
Sbjct: 33747 TGAFNALLKTLEEPTENVVFILATTELHKIPATILSR 33857
>Contig203
          Length = 23,545
  Plus Strand HSPs:
 Score = 100 (46.1 bits), Expect = 4.0e-07, P = 4.0e-07
 Identities = 22/71 (30%), Positives = 39/71 (54%), Frame = +1
Query:    31 DAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWFFLATREPER 90
             D V+E+    ++        +V  + D    +   AAN+LLK +EEP  E + FL T +  +
Sbjct: 18139 DVVKEMMANFSQTGYENKRQVFIIKDCDKMHINAANSLLKYIEEPQGEAYIFLLTNDDNK 18318
Query:    91 LLATLRSRCRL 101
             +L T++SR ++
Sbjct: 18319 VLPTIKSRTQV 18351
 Score = 58 (26.8 bits), Expect = 0.00034, Poisson P(2) = 0.00034
 Identities = 10/21 (47%), Positives = 12/21 (57%), Frame = +1
Query:     1 HCRGCQLMQAGTHPDYYTLAP 21
             HCR CQL++ G   D   L P
Sbjct: 18055 HCRSCQLIEQGDFADVTVLEP 18117
>Contig215
          Length = 27,361
  Plus Strand HSPs:
 Score = 49 (22.6 bits), Expect = 5.8, P = 1.0
 Identities = 11/30 (36%), Positives = 19/30 (63%), Frame = +3
Query:   168 DWYSLLAALNHEQAPARLHWLATLLMDALK 197
```

Following those BLAST hits is the sequence of the <u>contig</u> containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  delta prime
        (334 letters)

Database:  /usr/local/db/e_faecalis
           293 sequences; 3,209,119 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                   Smallest
                                                                     Sum
                                                    Reading  High  Probability
Sequences producing High-scoring Segment Pairs:       Frame Score  P(N)      N

6277                                                    -1    210  9.6e-16   1
6250                                                    -2    162  2.9e-10   1


>6277
     Length = 9336

  Minus Strand HSPs:

 Score = 210 (73.9 bits), Expect = 9.6e-16, P = 9.6e-16
 Identities = 62/218 (28%), Positives = 105/218 (48%), Frame = -1

Query:    11 FEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQA 70
             +++L  S++ GR  HA L +   G G        +++++ C       + C  C  C  +
Sbjct:  8865 YKQLQKSFEHGRLAHAYLFEGDTGTGKQEFGLWMAKHVFCTNLVNQQPCNECHNCVRINE 8686

Query:    71 GTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLK 130
               HPD    +AP+ G+ T+ V+ +RE+  + ++        KV  + +A  ++  AAN+LLK
Sbjct:  8685 NEHPDVLRIAPD-GQ-TIKVNQIRELKAEFSKSGVETAKKVFLIQEADKMSTGAANSLLK 8512

Query:   131 TLEEPPAETWFFLATREPERLLATLRSRCR-LHYLAGPPEQYAVTWLSREVTMSQDALLA 189
              LEEP  +      L T    R+L T++SRC+ LH+       +       + + +      LLA
Sbjct:  8511 FLEEPEGQILAILETTSLSRILPTIQSRCQTLHFQPLVKKTLIDRLIKQGIGEKTATLLA 8332

Query:   190 ALRLSAGSPGAALALFQGDNW--QARETLCQALAYSVPSGD 228
              L     S    A+ + Q D W  +ARE + Q    Y + S D
Sbjct:  8331 EL---TNSFEKAVEISQ-DEWFNEAREIILQWFNY-LKSND 8224


>6250
     Length = 24,587
```

```
Minus Strand HSPs:

 Score = 162 (57.0 bits), Expect = 2.9e-10, P = 2.9e-10
 Identities = 41/134 (30%), Positives = 62/134 (46%), Frame = -2

Query:    25 HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
             HA L      G G  +       ++ + C+  Q   + C  C  C  +  G   D   +   +
Sbjct:  5419 HAYLFTGPRGTGKTSAAKIFAKAINCKHSQDGEPCNVCETCVAITEGRLNDVIEI--DAA 5246

Query:    85 KNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWFFLA 144
              N  GV+ +R++ +K               KV   + +  +L+  A NALLKTLEEPP    F LA
Sbjct:  5245 SNN-GVEEIRDIRDKAKYAPTQAEYKVYIIDEVHMLSTGAFNALLKTLEEPPQNVIFILA 5069

Query:   145 TREPERLLATLRSR 158
             T EP ++   T+ SR
Sbjct:  5068 TTEPHKIPLTIISR 5027


Parameters:
  B=5

  ctxfactor=6.00
  E=10

  Query                              ----- As Used -----     ----- Computed ----
  Frame  MatID Matrix name          Lambda    K      H       Lambda    K       H
   +0      0   BLOSUM62             0.321   0.136  0.423    same     same    same
               Q=9,R=2              0.244   0.0300 0.180     n/a      n/a     n/a

  Query
  Frame  MatID  Length  Eff.Length    E     S W   T  X   E2      S2
   +0      0     334      334        10.   59 3  13 22  0.069    37
                                                    33  0.063    42


Statistics:

  Database:  /usr/local/db/e_faecalis
    Title:  /usr/local/db/e_faecalis
    Release date:  unknown
    Posted date:  12:53 PM EST Dec 11, 1998
    Format:  BLAST
  # of letters in database:  3,209,119
  # of sequences in database:  293
  # of database sequences satisfying E:  2
  No. of states in DFA:  540 (57 KB)
  Total size of DFA:  97 KB (128 KB)
  Time to generate neighborhood:  0.00u 0.01s 0.01t  Elapsed: 00:00:00
  No. of threads or processors used:  1
  Search cpu time:  2.07u 0.01s 2.08t  Elapsed: 00:00:02
  Total cpu time:  2.08u 0.03s 2.11t  Elapsed: 00:00:02
  Start:  Wed Mar 17 09:11:29 1999   End:  Wed Mar 17 09:11:31 1999
```

The top-scoring match came from this contig (up to 1000bp on either side of the hit are shown):

```
>6277 (from 7224 to 9336)
TTCAAACAACACATTAAGCGGCCACATAATCCCGAAATTTTGACAGGATTTAAAGATAAC
CCTTGATCTTTAGCCATTTTGATTGAAACTGGCATAAAATCTCCTAGAAATGTTGAGCAA
CATAGTTGTCTGCCACAAGGGCCAATGCCACCTAATATTTTCGCTTCATCTCGGACACCA
```

Following those BLAST hits is the sequence of the contig containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  deltaprime.ecoli
        (334 letters)

Database:  /usr/local/db/e_faecalis
           293 sequences; 3,209,119 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                   Smallest
                                                                     Sum
                                                          Reading  High  Probability
Sequences producing High-scoring Segment Pairs:             Frame Score  P(N)      N

6277                                                          -1   210  9.6e-16   1
6250                                                          -2   162  2.9e-10   1


>6277
     Length = 9336

  Minus Strand HSPs:

 Score = 210 (73.9 bits), Expect = 9.6e-16, P = 9.6e-16
 Identities = 62/218 (28%), Positives = 105/218 (48%), Frame = -1

Query:    11 FEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQA 70
             +++L  S++ GR  HA L +   G G       +++++ C      + C  C  C  +
Sbjct:  8865 YKQLQKSFEHGRLAHAYLFEGDTGTGKQEFGLWMAKHVFCTNLVNQQPCNECHNCVRINE 8686

Query:    71 GTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLK 130
               HPD   +AP+ G+ T+ V+ +RE+  + ++        KV  + +A  ++  AAN+LLK
Sbjct:  8685 NEHPDVLRIAPD-GQ-TIKVNQIRELKAEFSKSGVETAKKVFLIQEADKMSTGAANSLLK 8512

Query:   131 TLEEPPAETWFFLATREPERLLATLRSRCR-LHYLAGPPEQYAVTWLSREVTMSQDALLA 189
              LEEP  +    L T    R+L T++SRC+ LH+     +      + + +      LLA
Sbjct:  8511 FLEEPEGQILAILETTSLSRILPTIQSRCQTLHFQPLVKKTLIDRLIKQGIGEKTATLLA 8332

Query:   190 ALRLSAGSPGAALALFQGDNW--QARETLCQALAYSVPSGD 228
              L     S   A+ + Q D W  +ARE + Q    Y + S D
Sbjct:  8331 EL---TNSFEKAVEISQ-DEWFNEAREIILQWFNY-LKSND 8224

>6250
     Length = 24,587
```

```
Minus Strand HSPs:

 Score = 162 (57.0 bits), Expect = 2.9e-10, P = 2.9e-10
 Identities = 41/134 (30%), Positives = 62/134 (46%), Frame = -2

Query:    25 HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
             HA L      G G  +      ++ + C+  Q   + C  C  C  + G   D    +   +
Sbjct:  5419 HAYLFTGPRGTGKTSAAKIFAKAINCKHSQDGEPCNVCETCVAITEGRLNDVIEI--DAA 5246

Query:    85 KNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWFFLA 144
              N  GV+ +R++ +K              KV  + +   +L+  A NALLKTLEEPP     F LA
Sbjct:  5245 SNN-GVEEIRDIRDKAKYAPTQAEYKVYIIDEVHMLSTGAFNALLKTLEEPPQNVIFILA 5069

Query:   145 TREPERLLATLRSR 158
             T EP ++   T+ SR
Sbjct:  5068 TTEPHKIPLTIISR 5027


Parameters:
  B=5

  ctxfactor=6.00
  E=10

  Query                            -----  As Used  -----    -----  Computed   ----
  Frame  MatID Matrix name         Lambda    K        H       Lambda    K        H
   +0      0   BLOSUM62            0.321   0.136    0.423    same     same     same
               Q=9,R=2             0.244   0.0300   0.180     n/a      n/a      n/a

  Query
  Frame  MatID  Length  Eff.Length     E     S W    T  X   E2      S2
   +0      0     334       334        10.   59 3   13 22  0.069    37
                                                       33  0.063    42


Statistics:

  Database:  /usr/local/db/e_faecalis
    Title:  /usr/local/db/e_faecalis
    Release date:  unknown
    Posted date:  12:53 PM EST Dec 11, 1998
    Format:  BLAST
  # of letters in database:  3,209,119
  # of sequences in database:  293
  # of database sequences satisfying E:  2
  No. of states in DFA:  540 (57 KB)
  Total size of DFA:  97 KB (128 KB)
  Time to generate neighborhood:  0.00u 0.00s 0.00t  Elapsed: 00:00:00
  No. of threads or processors used:  1
  Search cpu time:  2.06u 0.02s 2.08t  Elapsed: 00:00:02
  Total cpu time:  2.08u 0.03s 2.11t  Elapsed: 00:00:02
  Start:  Wed Mar 17 10:15:00 1999   End:  Wed Mar 17 10:15:02 1999
```

The top-scoring match came from this contig (up to 1000bp on either side of the hit are shown):

```
>6277 (from 7224 to 9336)
TTCAAACAACACATTAAGCGGCCACATAATCCCGAAATTTTGACAGGATTTAAAGATAAC
CCTTGATCTTTAGCCATTTTGATTGAAACTGGCATAAAATCTCCTAGAAATGTTGAGCAA
CATAGTTGTCTGCCACAAGGGCCAATGCCACCTAATATTTTCGCTTCATCTCGGACACCA
```

ATTTGACGTAACTCAATTCGCGTCCGGAAAATAGCCGCTAAGTCTTTGACTAATTCACGA
AAATCAATTCGCCCATCTGCCGTAAAGTAAAAAATCATTTTGCTACGATCGAAGGTATAT
TCTACTCGCACTAATTTCATTTTTAAGTCATGAGCTCGAATTTTTTCATTGGCAATGCTT
TTGGCAGCTTCTGCATCAGCCAAATTTTTTTGTTCTTTTTCTAAATCATTGGCTGTTGCT
TTATTTAAAATGGGTTTTAGGTCCTCTGGTAAATCGTCTGAATCGACTGTTTTTTTAGGA
ATAGCAACAGTAGCTAATTGTTTTGACTGTTGAGATTCAACGAGTACTTTCTCATTATAA
ATATACTCAGATTTTCCAGGAGCAAAATAATAGATATGACCGGCTTCACGGAAGCGAACT
CCTACTACTTCTACCATTTTATTCCTCCTAATCTAGTTCAAGTGAACGTCGCTTCAATCG
GTTAAGGAACTTGCTGAAACAACTAAGTTCCTACTATATTATGAAACTGAATGCCACTTG
GCACTTTTTTCCTTTATGATTTAGGGTGAATCATTTGGATAACTAATTGTTCACAAACAT
TTTGCCAACTAACATTGGCAGTCCATTTTTGGCGTGCTTTCAAAATTAGCGCCAACCGTT
CCGCCTGCTCTTCCGTTACTTTTTTGGCTTGCTGTGTCGCAACACTTTCTTCCAATAATT
GACGGTAATAAACCATGAGCAAGTCAAAGCTAAGCGCTTGTTGTTCTTTTTCCTTAAATA
CTTTGACCATTTTCTTCTGAACGTAGATAAATGCCTGTAAATCATTACTTTTTAGATAAT
TAAACCATTGCAAAATGATTTCCCTAGCTTCATTAAACCATTCATCTTGAGAGATTTCAA
CTGCTTTCTCAAAACTATTTGTCAGTTCAGCTAAAAGGGTTGCAGTCTTTTCACCAATCC
CCTGTTTGATTAAGCGATCAATTAATGTTTTTTTGACTAATGGTTGAAAATGTAAGGTTT
GGCATCGTGATTGAATCGTTGGTAAAATTCGAGAAAGCGAAGTGGTTTCTAAAATAGCTA
AAATTTGTCCTTCTGGTTCTTCTAAAAATTTTAAGAGACTATTAGCTGCGCCGGTACTCA
TTTTATCTGCTTCTTGAATTAAGAAAACTTTTTTAGCAGTCTCGACCCCACTTTTAGAAA
ACTCCGCTTTTAATTCACGGATTTGGTTCACTTTGATGGTTTGCCCATCTGGCGCAATTC
TTAAAACATCTGGATGTTCATTTTCATTAATCCGCACACAATTATGGCATTCGTTACAAG
GCTGTTGATTTACTAAATTCGTACAAAAGACATGTTTCGCCATCCATAAGCCAAATTCTT
GTTTTCCAGTTCCTGTATCTCCTTCAAAAAGATAAGCATGGGCAAGACGACCATGCTCAA
AACTTTTTTGGAGTTGCTTGTACAGCAAAGGTTGCATTTGCTGTAGCTGTTGTGCTTCAT
TCATCTTAATATTGATGGAATCCTTCAACTGGTAAGACGAAGCAAGTAGCGCCGCCTACT
TCAACTTCCACAGGATAAGGAATTTGGCCATCCATTGTGATATCTAAAGTCACAGGTGTT
GAAACATATTGTTTTCTTGATTGACATGTTTCTTTAATTAAAGCTAATGTTTCGTCGACA
CGTTCATCATCAATCCCAATAATAAATGTGCTGTTTCCCGCTTTTAAGAACCCACCTGTT
GAGGATAATTTTGTAGCACGAATATTGGCATCAATAAATTCGTTGGCTAATCGGTTACTA
TCTTTGTCTTGTACAATGGCTAAAATAATCTTCATGGTCTACACCTTCCTATAATTAAAA
GTTTTCTGGATAACGTTCAATAATCGCCTGATACGTTGCTTCTACGACAAGTTCTAAACT
CATCCGTGCATCA

Following those BLAST hits is the sequence of the contig containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  delta prime
        (334 letters)

Database:  /usr/local/db/s_pneumoniae
           270 sequences; 2,114,666 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                     Smallest
                                                                       Sum
                                                    Reading  High  Probability
Sequences producing High-scoring Segment Pairs:       Frame Score  P(N)      N

sp_68                                                    -3   179  2.4e-12   1
sp_36                                                    +1   176  5.3e-12   1


>sp_68
      Length = 21,744

  Minus Strand HSPs:

 Score = 179 (63.0 bits), Expect = 2.4e-12, P = 2.4e-12
 Identities = 66/236 (27%), Positives = 109/236 (46%), Frame = -3

Query:    25 HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
             HA L      G G  ++     ++ + C     G + C +C  CQ +  G+  D     +   +
Sbjct: 17440 HAYLFSGPRGTGKTSVAKIFAKAMNCPNQVGGEPCNNCYICQAVTDGSLEDVIEM--DAA 17267

Query:    85 KNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEEPPAETWFFLA 144
              N  GVD +RE+ +K         L     KV  + +   +L+  A NALLKTLEEP       F LA
Sbjct: 17266 SNN-GVDEIREIRDKSTYAPSLARYKVYIIDEVHMLSTGAFNALLKTLEEPTQNVVFILA 17090

Query:   145 TREPERLLATLRSRC-RLHYLAGPPE---QYAVTWLSRE-VTMSQDAL-LAALRLSAGSP 198
             T E   ++ AT+ SR   R  + +     +    ++       L +E ++     +A+ + A R    G
Sbjct: 17089 TTELHKIPATILSRVQRFEFKSIKTQDIKEHIHYILEKENISSEPEAVEIIARRAEGGMR 16910

Query:   199 GA------ALALFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQAPARLHWLATLL 252
              A        AL+L QG+   +      + + +  ++             +AAL+  +    P  L   L  LL
Sbjct: 16909 DALSILDQALSLTQGN--ELTTAISEEITGTISLSALDDYVAALSQQDVPKALSCL-NLL 16739

Query:   253 MD 254
              D
Sbjct: 16738 FD 16733
```

```
>sp_36
      Length = 43,015

  Plus Strand HSPs:

 Score = 176 (62.0 bits), Expect = 5.3e-12, P = 5.3e-12
 Identities = 50/205 (24%), Positives = 89/205 (43%), Frame = +1

Query:      6 WLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGC 65
              W     F++ V    +   + +HA L              +     L++ L C     G     C    CR C
Sbjct: 23515 WQPAQFDRFVRILEQDQLNHAYLFSGF--FESLEMAQFLAKSLFCTDKVGVLPCEKCRSC 23688

Query:     66 QLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAA 125
              +L++ G   PD     + P          +   + +RE+  + ++           +V   +   A   +     AA
Sbjct: 23689 KLIEQGEFPDVTLIKPVN--QVIKTERIRELVGQFSQAGIESQQQVFIIEQADKMHPNAA 23862

Query:    126 NALLKTLEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMSQD 185
              N+LLK +EEP +E + F   T + E++L T+RSR ++ +       E+   +   L +     + +
Sbjct: 23863 NSLLKVIEEPQSEVYIFFLTSDEEKMLPTIRSRTQIFHFK-KQEEKLILLLEQMGLVKKK 24039

Query:    186 ALLAALRLSAGSPGAALALFQGDNW 210
              A L A + S       A       Q     W
Sbjct: 24040 ATLLA-KFSQSRAEAEKLANQASFW 24111


Parameters:
  B=5

  ctxfactor=6.00
  E=10

  Query                                 -----  As Used  -----      -----  Computed  ----
  Frame   MatID Matrix name             Lambda    K        H       Lambda    K        H
   +0       0   BLOSUM62                0.321   0.136    0.423    same     same     same
                Q=9,R=2                 0.244   0.0300   0.180     n/a      n/a      n/a

  Query
  Frame   MatID  Length  Eff.Length       E     S W   T  X    E2     S2
   +0       0     334       334          10.   57 3  13 22  0.069   37
                                                       33  0.063   42


Statistics:

  Database:  /usr/local/db/s_pneumoniae
    Title:  /usr/local/db/s_pneumoniae
    Release date:  unknown
    Posted date:  12:57 PM EST Dec 11, 1998
    Format:  BLAST
  # of letters in database:  2,114,666
  # of sequences in database:  270
  # of database sequences satisfying E:  2
  No. of states in DFA:  540 (57 KB)
  Total size of DFA:  97 KB (128 KB)
  Time to generate neighborhood:  0.00u 0.00s 0.00t  Elapsed:  00:00:00
  No. of threads or processors used:  1
  Search cpu time:  1.44u 0.01s 1.45t  Elapsed:  00:00:02
  Total cpu time:  1.45u 0.02s 1.47t  Elapsed:  00:00:02
  Start:  Wed Mar 17 09:13:52 1999   End:  Wed Mar 17 09:13:54 1999
```

The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*

Gerard Deckert*†, Patrick V. Warren*†, Terry Gaasterland‡, William G. Young*, Anna L. Lenox*, David E. Graham§, Ross Overbeek‡, Marjory A. Snead*, Martin Keller*, Monette Aujay*, Robert Huber‖, Robert A. Feldman*, Jay M. Short*, Gary J. Olsen§ & Ronald V. Swanson*

* *Diversa Corporation, 10665 Sorrento Valley Road, San Diego, California 92121, USA*
‡ *Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*
§ *Department of Microbiology, University of Illinois, Urbana, Illinois 61801, USA*
‖ *Lehrstuhl für Mikrobiologie, Universität Regensburg W-8400, Regensburg W-8400, Germany*

Aquifex aeolicus **was one of the earliest diverging, and is one of the most thermophilic, bacteria known. It can grow on hydrogen, oxygen, carbon dioxide, and mineral salts. The complex metabolic machinery needed for *A. aeolicus* to function as a chemolithoautotroph (an organism which uses an inorganic carbon source for biosynthesis and an inorganic chemical energy source) is encoded within a genome that is only one-third the size of the *E. coli* genome. Metabolic flexibility seems to be reduced as a result of the limited genome size. The use of oxygen (albeit at very low concentrations) as an electron acceptor is allowed by the presence of a complex respiratory apparatus. Although this organism grows at 95 °C, the extreme thermal limit of the Bacteria, only a few specific indications of thermophily are apparent from the genome. Here we describe the complete genome sequence of 1,551,335 base pairs of this evolutionarily and physiologically interesting organism.**

Complete genome sequences have been determined for a number of organisms, including Archaea[1], Bacteria[2–7], and Eukarya[8]. Here we present and explore the genome sequence of *Aquifex aeolicus*. With growth-temperature maxima near 95 °C, *Aquifex pyrophilus* and *A. aeolicus* are the most thermophilic bacteria known. Although isolated and described only recently[9], these species are related to filamentous bacteria first observed at the turn of the century, growing at 89 °C in the outflow of hot springs in Yellowstone National Park[10,11]. The observation of these macroscopic assemblages would later be instrumental in the drive to culture hyperthermophilic organisms[12].

The *Aquificaceae* represent the most deeply branching family within the bacterial domain on the basis of phylogenetic analysis of 16S ribosomal RNA sequences[13,14], although analyses of individual protein sequences vary in their placement of *Aquifex* relative to other groups[15–18]. The genera in this group, *Aquifex* and *Hydrogenobacter*, are thermophilic, hydrogen-oxidizing, microaerophilic, obligate chemolithoautotrophs[9,19–21]. *A. aeolicus* (isolated by R.H. and K. O. Stetter) was cultured at 85 °C under an $H_2/CO_2/O_2$ (79.5:19.5:1.0) atmosphere in a medium containing only inorganic components. *A. aeolicus* does not grow on a number of organic substrates, including sugars, amino acids, yeast extract or meat extract. Unlike its close relative *A. pyrophilus*, *A. aeolicus* has not been shown to grow anaerobically with nitrate as an electron acceptor in the laboratory.

From study of the physiology of the organism, several predictions can be made. As an autotroph, *A. aeolicus* must have genes encoding proteins for one or more modes of carbon fixation and a complete set of biosynthetic genes. As autotrophy is a feature that is distributed throughout the Archaea and Bacteria, most of the associated genes are expected to be of ancient origin and clearly related to those characterized elsewhere. The obligate autotrophy suggests a biosynthetic rather than a degradative character. Oxygen respiration implies the presence of corresponding utilization and tolerance genes. The early divergence of the *Aquificaceae* inferred from ribosomal RNA sequences leads to several questions. Are the machineries for oxygen usage and tolerance homologous to those found in mitochondria, and well studied organisms such as *Escherichia coli*, or were they invented separately? If there was far less oxygen when the lineage originated, is there evidence for use of alternative oxidants?

Genome

General features of the *A. aeolicus* genome are listed in Box 1. We classified 1,512 open-reading frames (ORFs) into one of three categories, namely, identified (Table 1), hypothetical, or unknown. Identified ORFs were further classified into one of 57 cellular role categories adapted from Riley[22] (Table 1). The relatively high G + C content of the two 16S-23S-5S rRNA operons (65%) is characteristic of thermophilic bacterial rRNAs[23]. The genome is densely packed: most genes are apparently expressed in polycistronic operons and many convergently transcribed genes overlap slightly. Nonetheless, many genes that are functionally grouped within operons in other organisms, such as the tryptophan or histidine biosynthesis pathways, are found dispersed throughout the *A. aeolicus* genome or appear in novel operons. Even when they encode subunits of the same enzyme, the genes are often separated on the chromosome (for example, *gltB* and *gltD*, the genes encoding the large and small subunits of glutamate synthase). Operon organization of genes for the biosynthesis of amino acids is found in both Archaea and Bacteria but it is not universal in either group. *A. aeolicus* is extreme in that no two amino acid biosynthetic genes are found in the same operon. In contrast, genes required for electron transport, hydrogenase subunits, transport systems, ribosomal subunits, and flagella are often in functionally related operons in *A. aeolicus* (Fig. 1). No introns or inteins (protein splicing elements) were detected in the genome.

A single extrachromosomal element (ECE) was identified during sequencing. Sequence redundancy for the total project was calculated to be 4.83. The ECE, however, is significantly over-represented

† Present addresses: Codex Bioinformatics Services, PO Box 90273, San Diego, California 92169, USA (G.D.); Department of Bioinformatics, SmithKline Beecham Pharmaceuticals, Collegeville, Philadelphia 19426, USA (P.V.W.)

relative to the chromosome; when calculated independently for the final assemblies, redundancies are 4.73 and 8.76 for the chromosome and for the ECE, respectively. The ECE therefore appears to be present at roughly twice the copy number of the chromosome. Although no ORFs on the ECE can be assigned a function with confidence, except for a transposase, two of the predicted proteins show similarity to hypothetical proteins in the *Methanococcus jannaschii* genome[1]. One ORF on the ECE is also present in two identical copies on the *A. aeolicus* chromosome, providing evidence of genetic exchange between the chromosome and the ECE.

Reductive tricarboxylic acid cycle

As an autotroph, *A. aeolicus* obtains all necessary carbon by fixing CO_2 from the environment. An assay for activity of the reductive tricarboxylic acid (TCA) cycle in *A. pyrophilus* cell extracts showed *in vitro* activities for each proposed reaction[24]. The reductive (reverse) TCA cycle fixes two molecules of CO_2 to form acetyl-coenzyme A (acetyl-CoA) and other biosynthetic intermediates[25]. The *A. aeolicus* genome contains genes encoding malate dehydrogenase, fumarate hydratase, fumarate reductase, succinate-CoA ligase, ferredoxin oxidoreductase, isocitrate dehydrogenase, aconitase and citrate synthase, which together could constitute the TCA pathway. There is no biochemical evidence for alternative carbon-fixation pathways in *A. pyrophilus*[24,25] nor is there sequence evidence for such pathways in *A. aeolicus.*

The TCA cycle is vital as it provides the substrates of many biosynthetic pathways. (It is beyond the scope of this report to detail these biosynthetic pathways, but they seem to be typically bacterial, and candidate genes for all or most of the enzymes have been identified in *A. aeolicus.*) The central role of the TCA cycle is emphasized by duplication of many of its constituent genes in *A. aeolicus.* Two genes encode proteins that are similar to malate dehydrogenase (in addition to a lactate dehydrogenase). The fumarate hydratase is split into amino- and carboxy-terminal subunits, as is the case in *M. jannaschii*[1]. Unlinked genes encoding two iron–sulphur proteins of fumarate reductase (alternatively succinate dehydrogenase) accompany a single flavoprotein subunit. Two sets of genes resembling succinate-CoA ligase (both the α- and β-subunits) are present. *A. aeolicus* has two putative operons encoding four-subunit (α, β, γ, δ) 2-acid ferredoxin oxidoreductases; members of this family catalyze reversible carboxylation/decarboxylation of pyruvate, 2-isoketovalerate, or 2-oxoglutarate with varying specificity[26]. These duplicated genes may encode paralogous proteins with unique substrate specificity, as opposed to redundant functions. For example, a paralogue of succinate-CoA ligase may activate citrate with coenzyme A to form citryl-CoA, which citrate synthase can cleave to produce oxaloacetate and acetyl-CoA.

Gluconeogenesis through the Embden–Meyerhof–Parnas pathway

Growing autotrophically, *A. aeolicus* must synthesize pentose and hexose monosaccharides from products of the reductive TCA cycle. Pyruvate produced by pyruvate ferredoxin oxidoreductase or by pyruvate carboxylase (oxaloacetate decarboxylase)[24] may enter the Embden–Meyerhof–Parnas pathway of glycolysis and gluconeogenesis. Genes encoding fructose-1,6-bisphosphatase, an essential gluconeogenic enzyme in *E. coli*, have not been identified in the genomes of the autotrophs *A. aeolicus* or *M. jannaschii*[1], suggesting that an unidentified pathway may exist. The *A. aeolicus* genome also encodes enzymes of the pentose-phosphate pathway and enzymes for glycogen synthesis and catabolism. We found neither (phospho) gluconate dehydrase nor 2-keto-3-deoxy-(6-phospho)gluconate aldolase of the Entner–Doudoroff pathway.

Respiration

Aquifex species are able to grow by using oxygen concentrations as low as 7.5 p.p.m. (R.H. and K. O. Stetter, unpublished observations). The enzymes for oxygen respiration are similar to those of other bacteria: ubiquinol cytochrome *c* oxidoreductase (*bc*$_1$ complex), cytochrome *c* (three different genes) and cytochrome *c* oxidase (with two different subunit I genes and two different subunit II genes). The alternative system, with cytochrome *bd* ubiquinol oxidase, is also present. Clearly, the *Aquifex* lineage did not independently invent oxygen respiration. This leaves at least three possibilities: consistent with the ability of *Aquifex* to use very low levels of oxygen, the oxygen-respiration system was highly developed when oxygen had only a small fraction of its present concentration before the advent of oxygenic photosynthesis; contrary to what is implied by the 16S phylogeny, the lineage including *Aquifex* originated after the rise in atmospheric oxygen; or oxygen respiration developed once, and was then laterally transferred among bacterial lineages and acquired by *Aquifex.*

Many other oxidoreductases are present in addition to those obviously involved in oxygen respiration. The physiological role of most of these oxidoreductases is unknown or ambiguous, but two deserve comment. There is a putative nitrate reductase in the genome, although *A. aeolicus* has not been observed to perform NO_3^- respiration, unlike the closely related *A. pyrophilus.* The nitrate reductase gene is adjacent to a nitrate transporter, and may be involved in nitrogen assimilation rather than respiration. It is also possible that *A. aeolicus* has a latent ability to respire with nitrate but that the conditions required have not been found. Two gene sequences show strong similarities to Rieske proteins, even though the rest of the ubiquinol cytochrome *c* oxidoreductase subunits appear only once in the genome. One of these Rieske protein genes is adjacent to a sulphide dehydrogenase subunit, suggesting a role in sulphur respiration.

Oxidative stress

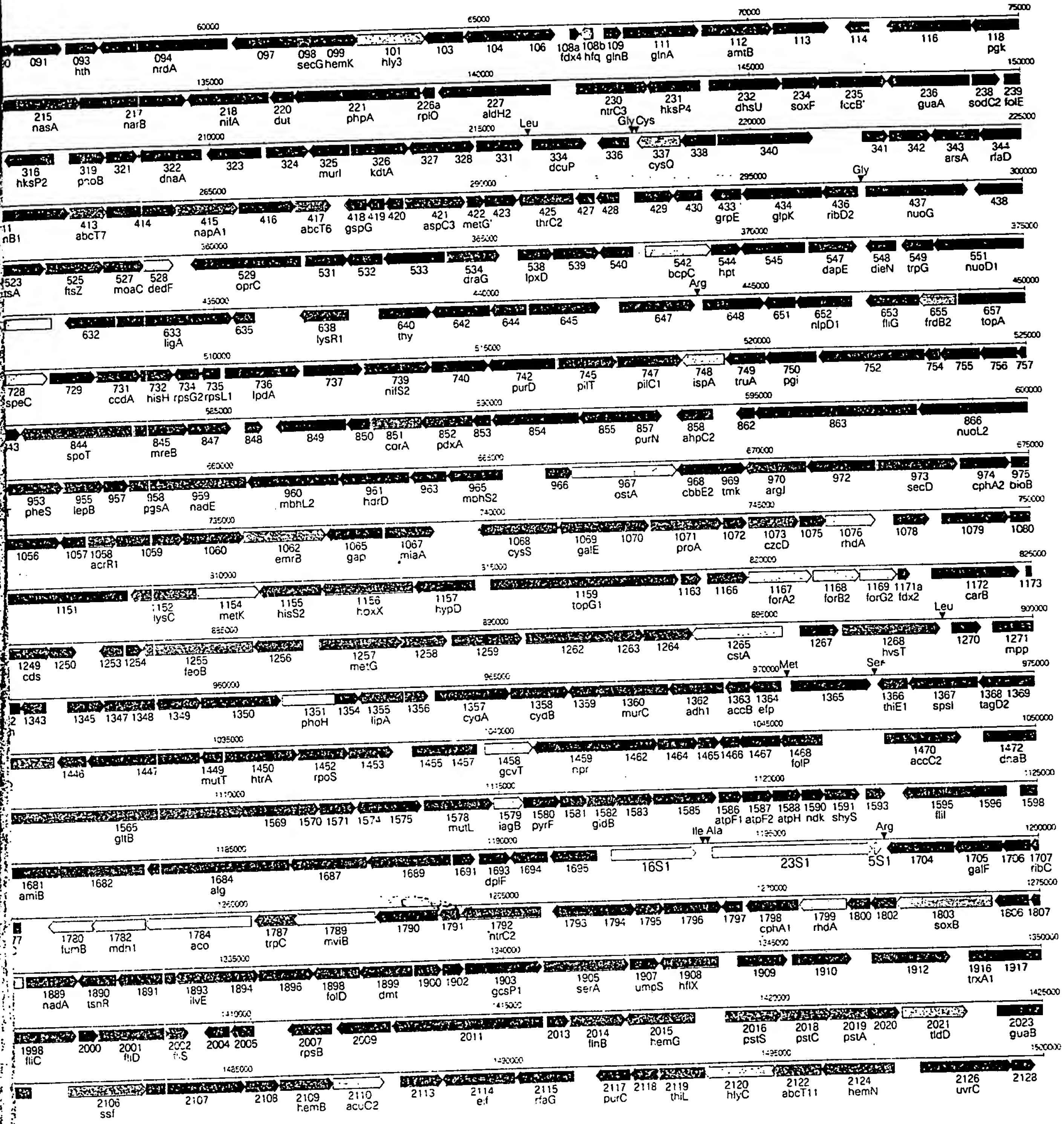
A. aeolicus grows optimally under microaerophilic conditions and consequently possesses various protective enzymes to counter reactive oxygen species, particularly superoxide and peroxide. The genome contains three genes encoding superoxide dismutases, two of the copper/zinc family and one of the iron/manganese family. The latter has also been noted in *A. pyrophilus*[27]. One of the copper/zinc superoxide dismutase genes is located in a large gene cluster encoding formate dehydrogenase.

No catalase genes were identified. There are several genes in the genome that might encode proteins that catalyze the detoxification of H_2O_2, including cytochrome *c* peroxidase, thiol peroxidase, and two alkyl hydroperoxide reductase genes. All of these enzymes require an exogenous reductant and therefore do not evolve O_2. However, treatment of *A. pyrophilus*[9] or *A. aeolicus* biomass with H_2O_2 results in the rapid evolution of gas bubbles. This catalase activity may result from a novel enzyme that cannot yet be identified by sequence similarity.

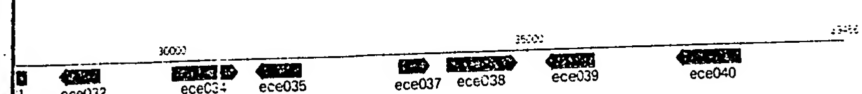
Motility

Like *A. pyrophilus*[9], *A. aeolicus* is motile and possesses monopolar polytrichous flagella. More than 25 genes encoding proteins involved in flagellar structure and biosynthesis have been identified in *A. aeolicus* (Box 1). However, no homologues of the bacterial chemotaxis system were identified. In enteric bacteria, membrane-bound receptors bind chemoattractants and repellents and mod-

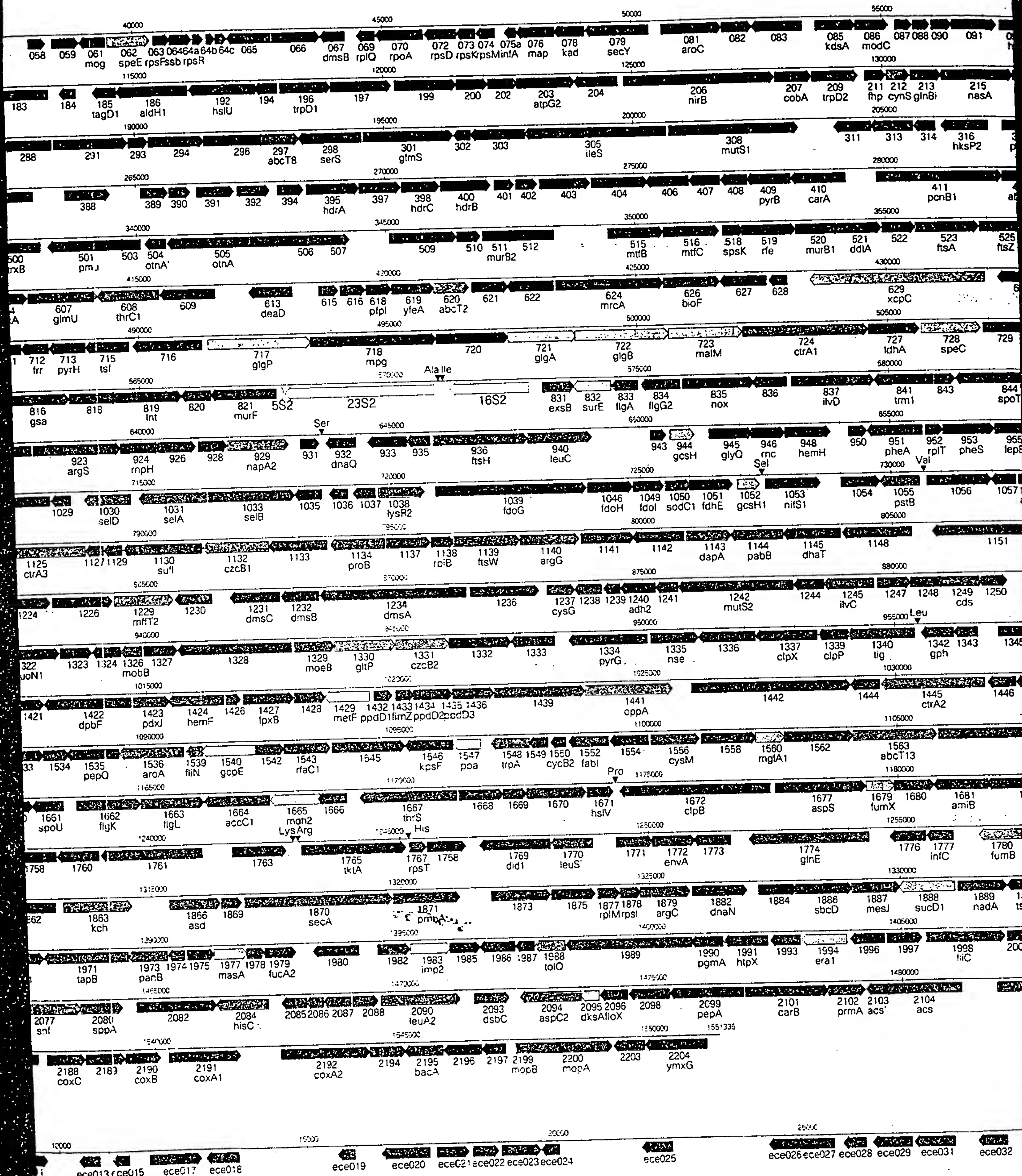
Figure 1 Linear map of the *A. aeolicus* circular chromosome. Genes are shown as arrows which denote the direction of transcription and are coloured to denote functional categorization according to the key below the figure. The sequences of the two rRNA gene clusters are identical. Here, the first base of the coding sequence of *fusA* was arbitrarily assigned as base number 1 as no origin of replication has been identified. ORF numbers are discontinuous because some ORFs representing 100 amino acids or more are not predicted to be coding and are not shown.



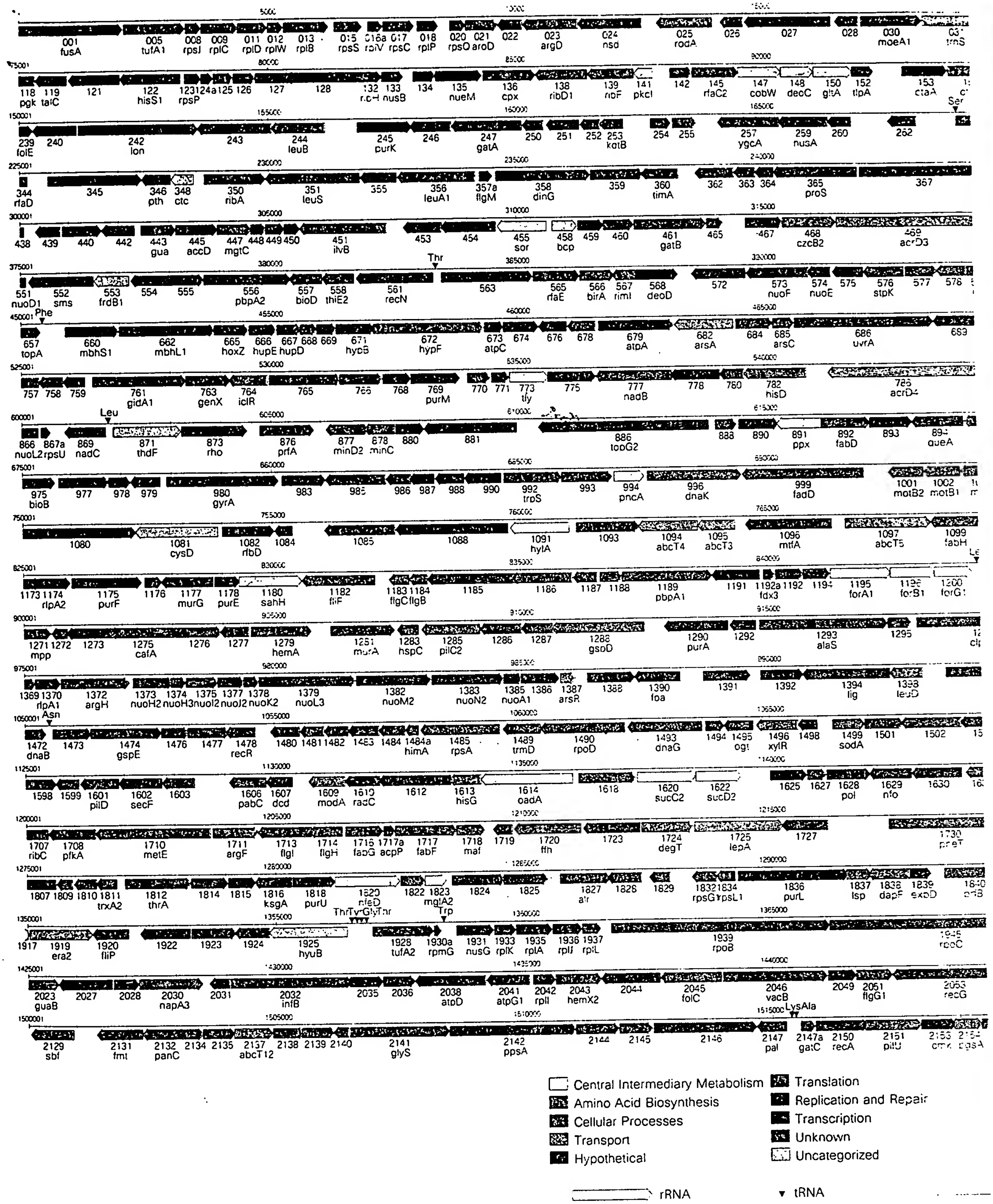
60000 65000 70000 75000
091 093 hth 094 nrdA 097 098 secG 099 hemK 101 hly3 103 104 106 108a fdx4 108b hfq 109 glnB 111 glnA 112 amtB 113 114 116 118 pgk
135000 140000 145000 150000
215 nasA 217 narB 218 nilA 220 dut 221 phpA 226a rplO 227 aldH2 230 ntrC3 231 hksP4 232 dhsU 234 soxF 235 fccB' 236 guaA 238 sodC2 239 folE
210000 215000 Leu Gly Cys 220000 225000
316 hksP2 319 proB 321 322 dnaA 323 324 325 murI 326 kdtA 327 328 331 334 dcuP 336 337 cysQ 338 340 341 342 343 arsA 344 rfaD
Gly 295000 300000
265000 290000
nB1 413 abcT7 414 415 napA1 416 417 abcT6 418 419 420 gspG 421 aspC3 422 metG' 423 425 thrC2 427 428 429 430 433 grpE 434 glpK 436 ribD2 437 nuoG 438
360000 365000 370000 375000
ftsA 525 ftsZ 527 moaC 528 dedF 529 oprC 531 532 533 534 draG 538 lpxD 539 540 542 bcpC 544 hpt 545 547 dapE 548 dieN 549 trpG 551 nuoD1
Arg 435000 440000 445000 450000
632 633 ligA 635 638 lysR1 640 thy 642 644 645 647 648 651 652 nlpD1 653 fliG 655 frdB2 657 topA
510000 515000 520000 525000
728 speC 729 731 ccdA 732 hisH 734 rpsG2 735 rpsL1 736 lpdA 737 739 nifS2 740 742 purD 745 pilT 747 pilC1 748 ispA 749 truA 750 pgi 752 754 755 756 757
585000 590000 595000 600000
843 844 spoT 845 mreB 847 848 849 850 851 corA 852 pdxA 853 854 855 857 purN 858 ahpC2 862 863 866 nuoL2
660000 665000 670000 675000
953 pheS 955 lepB 957 958 pgsA 959 nadE 960 mbhL2 961 hdrD 963 965 mohS2 966 967 ostA 968 cbbE2 969 tmk 970 argJ 972 973 secD 974 cphA2 975 bioB
735000 740000 745000 750000
1056 1057 1058 acrR1 1059 1060 1062 emrB 1065 gap 1067 miaA 1068 cysS 1069 galE 1070 1071 proA 1072 1073 czcD 1075 1076 rhdA 1078 1079 1080
810000 815000 820000 825000
1151 1152 lysC 1154 metK 1155 hisS2 1156 hoxX 1157 hypD 1159 topG1 1163 1166 1167 forA2 1168 forB2 1169 forG2 1171a fdx2 1172 carB 1173
885000 890000 895000 Leu 900000
1249 cds 1250 1253 1254 1255 feoB 1256 1257 metG 1258 1259 1262 1263 1264 1265 cstA 1267 1268 hvsT 1270 1271 mpp
960000 965000 970000 Met Ser 975000
1343 1345 1347 1348 1349 1350 1351 phoH 1354 1355 lipA 1356 1357 cydA 1358 cydB 1359 1360 murC 1362 adh1 1363 accB 1364 efp 1365 1366 thiE1 1367 spsI 1368 tagD2 1369
1035000 1040000 1045000 1050000
1446 1447 1449 mutT 1450 htrA 1452 rpoS 1453 1455 1457 1458 gcvT 1459 npr 1462 1464 1465 1466 1467 1468 folP 1470 accC2 1472 dnaB
1110000 1115000 1120000 1125000
1565 gltB 1569 1570 1571 1574 1575 1578 mutL 1579 iagB 1580 pyrF 1581 1582 gidB 1583 1585 1586 atpF1 1587 atpF2 1588 atpH 1590 ndk 1591 shyS 1593 1595 fliI 1596 1598
Ile Ala Arg
1185000 1190000 1195000 1200000
1681 amiB 1682 1684 alg 1687 1689 1691 1693 dplF 1694 1695 16S1 23S1 5S1 1704 1705 galF 1706 1707 ribC
1260000 1265000 1270000 1275000
1780 fumB 1782 mdn1 1784 aco 1787 trpC 1789 mviB 1790 1791 1792 ntrC2 1793 1794 1795 1796 1797 1798 cphA1 1799 rhdA 1800 1802 1803 soxB 1806 1807
1335000 1340000 1345000 1350000
1889 nadA 1890 tsnR 1891 1893 ilvE 1894 1896 1898 folD 1899 dmt 1900 1902 1903 gcsP1 1905 serA 1907 umpS 1908 hflX 1909 1910 1912 1916 trxA1 1917
1410000 1415000 1420000 1425000
1998 fliC 2000 2001 fliD 2002 2004 2005 2007 rpsB 2009 2011 2013 2014 flnB 2015 hemG 2016 pstS 2018 pstC 2019 pstA 2020 2021 tldD 2023 guaB
1485000 1490000 1495000 1500000
2106 ssf 2107 2108 2109 hemB 2110 accC2 2113 2114 eif 2115 rfaG 2117 purC 2118 2119 thiL 2120 hlyC 2122 abcT11 2124 hemN 2126 uvrC 2128



30000 35000
ece032 ece034 ece035 ece037 ece038 ece039 ece040



40000 45000 50000 55000
058 059 061 mog 062 speE 063 rpsF 064 ssb 64a rpsR 64b 64c 065 066 067 dmsB 069 rplQ 070 rpoA 072 rpsD 073 rpsK 074 rpsM 075a infA 076 map 078 kad 079 secY 081 aroC 082 083 085 kdsA 086 modC 087 088 090 091
115000 120000 125000 130000
183 184 185 tagD1 186 aldH1 192 hslU 194 196 trpD1 197 199 200 202 203 atpG2 204 206 nirB 207 cobA 209 trpD2 211 fhp 212 cynS 213 glnBi 215 nasA
190000 195000 200000 205000
288 291 293 294 296 297 abcT8 298 serS 301 glmS 302 303 305 ileS 308 mutS1 311 313 314 316 hksP2
265000 270000 275000 280000
388 389 390 391 392 394 395 hdrA 397 398 hdrC 400 hdrB 401 402 403 404 406 407 408 409 pyrB 410 carA 411 pcnB1
340000 345000 350000 355000
500 trxB 501 prnJ 503 504 otnA' 505 otnA 506 507 509 510 511 murB2 512 515 mtfB 516 mtfC 518 spsK 519 rfe 520 murB1 521 ddlA 522 523 ftsA 525 ftsZ
415000 420000 425000 430000
607 glmU 608 thrC1 609 613 deaD 615 616 618 pfpI 619 yfeA 620 abcT2 621 622 624 mrcA 626 bioF 627 628 629 xcpC
490000 495000 500000 505000
712 frr 713 pyrH 715 tsf 716 717 glgP 718 mpg 720 721 glgA 722 glgB 723 malM 724 ctrA1 727 ldhA 728 speC 729
565000 570000 575000 580000
Ala Ile
816 gsa 818 819 int 820 821 murF 5S2 23S2 16S2 831 exsB 832 surE 833 flgA 834 flgG2 835 nox 836 837 ilvD 841 trm1 843 844 spoT
640000 645000 650000 655000
Ser
923 argS 924 rnpH 926 928 929 napA2 931 932 dnaQ 933 935 936 ftsH 940 leuC 943 944 gcsH 945 glyQ 946 rnc 948 hemH 950 951 pheA 952 rplT 953 pheS 955 lepB
Sel Val
715000 720000 725000 730000
1029 1030 selD 1031 selA 1033 selB 1035 1036 1037 1038 lysR2 1039 fdoG 1046 fdoH 1049 fdoI 1050 sodC1 1051 fdhE 1052 gcsH1 1053 nifS1 1054 1055 pstB 1056 1057
790000 795000 800000 805000
1125 ctrA3 1127 1129 1130 sufI 1132 czcB1 1133 1134 proB 1137 1138 rpiB 1139 ftsW 1140 argG 1141 1142 1143 dapA 1144 pabB 1145 dhaT 1148 1151
865000 870000 875000 880000
1224 1226 1229 mffT2 1230 1231 dmsC 1232 dmsB 1234 dmsA 1236 1237 cysG 1238 1239 1240 adh2 1241 1242 mutS2 1244 1245 ilvC 1247 1248 1249 cds 1250
940000 945000 950000 955000 Leu
1322 nuoN1 1323 1324 1326 mobB 1327 1328 1329 moeB 1330 gltP 1331 czcB2 1332 1333 1334 pyrG 1335 nse 1336 1337 clpX 1339 clpP 1340 tig 1342 gph 1343 1345
1015000 1020000 1025000 1030000
1421 1422 dpbF 1423 pdxJ 1424 hemF 1426 1427 lpxB 1428 1429 metF 1432 ppdD1 1433 fimZ 1434 ppdD2 1435 1436 pcdD3 1439 1441 oppA 1442 1444 1445 ctrA2 1446
1090000 1095000 1100000 1105000
1533 1534 1535 pepQ 1536 aroA 1539 fliN 1540 gcpE 1542 1543 rfaC1 1545 1546 kpsF 1547 poa 1548 trpA 1549 1550 cycB2 1552 fabI 1554 1556 cysM 1558 1560 mglA1 1562 1563 abcT13
Pro
1165000 1170000 1175000 1180000
1661 spoU 1662 flgK 1663 flgL 1664 accC1 1665 mdh2 1666 1667 thrS 1668 1669 1670 1671 hslV 1672 clpB 1677 aspS 1679 fumX 1680 1681 amiB
Lys Arg His
1240000 1245000 1250000 1255000
1758 1760 1761 1763 1765 tktA 1767 rpsT 1768 1769 did1 1770 leuS 1771 1772 envA 1773 1774 glnE 1776 1777 infC 1780 fumB
1315000 1320000 1325000 1330000
1862 1863 kch 1866 asd 1869 1870 secA 1871 pmbA 1873 1875 1877 rplM 1878 rpsI 1879 argC 1882 dnaN 1884 1886 sbcD 1887 mesJ 1888 sucD1 1889 nadA tsn
1390000 1395000 1400000 1405000
1971 tapB 1973 panB 1974 1975 1977 masA 1978 1979 fucA2 1980 1982 1983 imp2 1985 1986 1987 1988 tolQ 1989 1990 pgmA 1991 htpX 1993 1994 era1 1996 1997 1998 fliC 2000
1465000 1470000 1475000 1480000
2077 snf 2080 sppA 2082 2084 hisC 2085 2086 2087 2088 2090 leuA2 2093 dsbC 2094 aspC2 2095 2096 dksA floX 2098 2099 pepA 2101 carB 2102 prmA 2103 acs' 2104 acs
1540000 1545000 1550000 1551335
2188 coxC 2189 2190 coxB 2191 coxA1 2192 coxA2 2194 2195 bacA 2196 2197 2199 mopB 2200 mopA 2203 2204 ymxG
10000 15000 20000 25000
ece013 ece015 ece017 ece018 ece019 ece020 ece021 ece022 ece023 ece024 ece025 ece026 ece027 ece028 ece029 ece031 ece032



Central Intermediary Metabolism
Amino Acid Biosynthesis
Cellular Processes
Transport
Hypothetical
Translation
Replication and Repair
Transcription
Unknown
Uncategorized
rRNA
tRNA

Table 1 ***Aquifex aeolicus*** **Open Reading Frame Identifications. Gene numbers (Aq) correspond to those in Fig.1. Percentages refer to the identity found in the best FASTA alignment. The percentage of the sequence covered by the alignment is displayed with bullets as follows 20–40% · , 40–60% · · , 60–80% · · · , 80–100% · · · ·**

Amino Acid Biosynthesis			
Aromatic amino acids			
Aq1536	aroA	5-enolpyruvylshikimate-3-phosphate synthetase	43.0% ....
Aq081	aroC	chorismate synthase	55.2% ....
Aq021	aroD	3-dehydroquinate dehydratase	33.3% ....
Aq901	aroE	shikimate 5-dehydrogenase	46.1% ....
Aq2177	aroK	shikimate kinase	36.5% ....
Aq951	pheA	chorismate mutase/prephenate dehydratase	44.0% ....
Aq1548	trpA	tryptophan synthase alpha subunit	44.5% ....
Aq706	trpB1	tryptophan synthase beta subunit	68.0% ....
Aq1410	trpB2	tryptophan synthase beta subunit	50.0% ....
Aq1787	trpC	indole-3-glycerol phosphate synthase	43.3% ....
Aq196	trpD1	phosphoribosylanthranilate transferase	45.1% ....
Aq209	trpD2	phosphoribosylanthranilate transferase	24.9% ....
Aq582	trpE	anthranilate synthase component I	50.0% ....
Aq2076	trpF	phosphoribosyl anthranilate isomerase	45.6% ....
Aq549	trpG	anthranilate synthase component II	59.2% ....
Aq1755	tyrA	prephenate dehydrogenase	36.1% ....
Aspartate family			
Aq1866	asd	aspartate-semialdehyde dehydrogenase	54.6% ....
Aq1969	aspC1	aspartate aminotransferase	53.5% ....
Aq2094	aspC2	aminotransferase (AspC family)	55.4% ....
Aq421	aspC3	aminotransferase (AspC family)	43.3% ....
Aq273	aspC4	aminotransferase (AspC family)	48.5% ....
Aq1143	dapA	dihydrodipicolinate synthase	53.1% ....
Aq916	dapB	dihydrodipicolinate reductase	44.2% ....
Aq547	dapE	succinyl-diaminopimelate desuccinylase	25.8% ....
Aq1838	dapF	diaminopimelate epimerase	35.5% ....
Aq1208	lysA	diaminopimelate decarboxylase	47.4% ....
Aq1152	lysC	aspartokinase	52.2% ....
Aq1710	metE	tetrahydropteroyltriglutamate methyltransferase	45.9% ....
Aq1812	thrA	homoserine dehydrogenase	40.4% ....
Aq1309	thrB	homoserine kinase	38.3% ....
Aq608	thrC1	threonine synthase	64.3% ....
Aq425	thrC2	threonine synthase	61.9% ....
Branched-chain family			
Aq451	ilvB	acetolactate synthase large subunit	53.1% ....
Aq1245	ilvC	acetohydroxy acid isomeroreductase	64.3% ....
Aq837	ilvD	dihydroxyacid dehydratase	58.0% ....
Aq1893	ilvE	branched-chain amino acid aminotransferase	40.3% ....
Aq1851	ilvH	acetolactate synthase	53.2% ....
Aq356	leuA1	2-isopropylmalate synthase	52.1% ....
Aq2090	leuA2	2-isopropylmalate synthase	49.9% ....
Aq244	leuB	3-isopropylmalate dehydrogenase	58.7% ....
Aq940	leuC	large subunit of isopropylmalate isomerase	52.3% ....
Aq1398	leuD	3-isopropylmalate dehydratase	56.6% ...
Glutamate family			
Aq2068	argB	acetylglutamate kinase	54.2% ....
Aq1879	argC	N-Acetyl-gamma-glutamylphosphate reductase	40.6% ....
Aq023	argD	N-acetylornithine aminotransferase	49.5% ....
Aq1711	argF	ornithine carbamoyltransferase	46.2% ....
Aq1140	argG	argininosuccinate synthase	54.9% ....
Aq1372	argH	argininosuccinate lyase	46.4% ....
Aq970	argJ	glutamate N-acetyltransferase	39.8% ....
Aq111	glnA	glutamine synthetase	57.6% ....
Aq109	glnB	nitrogen regulatory PII protein	73.2% ....
Aq1774	glnE	glutamate ammonia ligase adenylyl-transferase	28.4% ....
Aq1565	gltB	glutamate synthase large subunit	44.3% ....
Aq2064	gltD	glutamate synthase small subunit gltD	37.7% ....
Aq1071	proA	gamma-glutamyl phosphate reductase	47.9% ....
Aq1134	proB	glutamate 5-kinase	43.2% ....
Aq166	proC	pyrroline carboxylate reductase	35.1% ....
Histidine			
Aq1303	hisA	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase	40.9% ....
Aq039	hisB	imidazoleglycerolphosphate dehydratase	46.4% ....
Aq2084	hisC	histidinol-phosphate aminotransferase	33.7% ....
Aq782	hisD	histidinol dehydrogenase	49.9% ....
Aq181	hisF	HisF (cyclase)	59.9% ....
Aq1613	hisG	ATP phosphoribosyltransferase	40.3% ....
Aq732	hisH	amidotransferase HisH	47.7% ....
Aq1968	hisIE	phosphoribosyl-ATP pyrophosphohydrolase	43.8% ....
Selenocysteine			
Aq1031	selA	L-seryl-tRNA(ser) selenium transferase	42.7% ....
Aq1030	selD	selenophosphate synthase	37.7% ....
Serine family			
Aq1556	cysM	cysteine synthase, O-acetylserine (thiol) lyase B	45.8% ....
Aq479	glyA	serine hydroxymethyl transferase	62.7% ....
Aq1905	serA	D-3-phosphoglycerate dehydrogenase	44.1% ....
Cell Envelope			
Pili and fimbrae			
Aq1433	fimZ	minor pilin	34.9% ..
Aq1432	ppdD1	pilin	40.6% ....
Aq1434	ppdD2	pilin	26.4% ....
Aq1435	ppdD3	pilin	28.2% ...
Lipoproteins and porins			
Aq270	lgt	prolipoprotein diacylglyceryl transferase	30.1% ....
Aq819	lnt	apolipoprotein N-acyltransferase	25.5% ....
Aq652	nlpD1	lipoprotein	25.4% ....
Aq1753	nlpD2	lipoprotein NlpD fragment	43.2% ....
Aq529	oprC	outer membrane protein c	27.2% ....
Aq2147	pal	peptidoglycan associated lipoprotein	35.1% ..
Aq1370	rlpA1	rare lipoprotein A	61.1% ..
Aq1174	rlpA2	rare lipoprotein A	40.6% ...
Aq2166	scbA	adhesion protein	25.7% ....
Aq619	yfeA	adhesion B precursor	28.5% ....
Peptidoglycan			
Aq1827	alr	alanine racemase	33.2% ....
Aq1681	amiB	N-acetylmuramoyl-L-alanine amidase	31.0% ....
Aq2195	bacA	undecaprenol kinase	43.1% ....
Aq1798	cphA1	beta lactamase precursor	25.0% ...
Aq974	cphA2	beta lactamase precursor	29.4% ...
Aq521	ddlA	D-alanine:D-alanine ligase	38.2% ....
Aq301	glmS	glucosamine-fructose-6-phosphate aminotransferase	43.2% ....
Aq607	glmU	UDP-N-acetylglucosamine pyrophosphorylase	37.6% ...
Aq053	mraY	phospho-N-acetylmuramoyl-pentapeptide-transferase	47.5% ....
Aq624	mrcA	penicillin binding protein 1A	33.2% ....
Aq1281	murA	UDP-N-acetylglucosamine	
		1-carboxyvinyltransferase	45.7% ....
Aq520	murB1	UDP-N-acetylenolpyruvoylglucosamine reductase	35.6% ....
Aq511	murB2	UDP-N-acetylenolpyruvoylglucosamine reductase	38.9% ....
Aq1360	murC	UDP-N-acetylmuramate-alanine ligase	46.1% ....
Aq2075	murD	UDP-N-acetylmuramoylalanine-D-glutamate ligase	29.3% ....
Aq1747	murE	UDP-MurNac-tripeptide synthetase	42.9% ....
Aq821	murF	UDP-MURNAC-pentapeptide sythetase	32.3% ....
Aq1177	murG	phospho-N-acetylmuramoyl-pentapeptide-transferase	30.5% ....
Aq325	murI	glutamate racemase	43.4% ....
Aq1189	pbpA1	penicillin binding protein 2	32.2% ....
Aq556	pbpA2	penicillin binding protein 2	30.3% ....
Aq185	tagD1	glycerol-3-phosphate cytidyltransferase	52.0% ....
Aq1368	tagD2	glycerol-3-phosphate cytidyltransferase	67.2% ...
Surface polysaccharides and lipopolysaccharides			
Aq1684	alg	alginate synthesis-related protein	37.2% ..
Aq1641	cap	capsular polysaccharide biosynthesis protein	30.8% ....
Aq1899	dmt	dolichol-phosphate mannosyltransferase	40.2% ....
Aq1772	envA	UDP-3-0-acyl N-acetylglcosamine deacetylase	36.5% ....
Aq1757	exbB	biopolymer transport exbB	48.2% ....
Aq1839	exbD	biopolymer transport ExbD	34.7% ....
Aq1069	galE	UDP-glucose-4-epimerase	54.7% ....
Aq1705	galF	UDP-glucose pyrophosphorylase	47.2% ....
Aq908	gmhA	phosphoheptose isomerase	63.4% ...
Aq085	kdsA	3-deoxy-d-manno-octulosonic acid 8-phosphate synthase	52.0% ....
Aq326	kdtA	3-deoxy-D-manno-2-octulosonic acid transferase	28.9% ....
Aq253	kdtB	lipopolysaccharide core biosynthesis protein	46.5% ....
Aq1546	kpsF	polysialic acid capsule expression protein	45.9% ....
Aq692	kpsU	3-deoxy-manno-octulosonate cytidylyltransferase	41.3% ....
Aq1742	lgtF	beta 1,4 glucosyltransferase	35.2% ....
Aq604	lpxA	acyl-[acyl-carrier-protein]-UDP-N-acetylglucosamine acyltransferase	47.7% ....
Aq1427	lpxB	lipid A disaccharide synthetase	31.6% ....
Aq538	lpxD	UDP-3-O-[3-hydroxymyristoyl] glucosamine N acyltransferase	43.3% ....
Aq718	mpg	mannose-1-phosphate guanyltransferase	34.1% ....
Aq1096	mtfA	mannosyltransferase A	34.3% ...
Aq515	mtfB	mannosyltransferase B	29.0% ....
Aq516	mtfC	mannosyltransferase C	35.9% ....
Aq1335	nse	nucleotide sugar epimerase	45.8% ....
Aq505	otnA	polysaccharide biosynthesis protein	26.9% ....
Aq504	otnA'	polysaccharide biosynthesis protein (fragment)	37.8% ..
Aq1543	rfaC1	ADP-heptose:LPS heptosyltransferase	30.7% ....
Aq145	rfaC2	ADP-heptose:LPS heptosyltransferase	28.1% ....
Aq344	rfaD	ADP-L-glycero-D-manno-heptose-6-epimerase	39.6% ....
Aq565	rfaE	ADP-heptose synthase	44.0% ....
Aq2115	rfaG	glucosyl transferase I	27.1% ....
Aq1082	rfbD	GDP-D-mannose dehydratase	53.2% ....
Aq519	rfe	undecaprenyl-phosphate-alpha-N-acetylglucosaminyltransferase	24.8% ....
Aq1367	spsI	glucose-1-phosphate thymidylyltransferase	30.4% ..
Aq518	spsK	spore coat polysaccharide biosynthesis protein SpsK	49.5% ...
Aq589	xanB	mannose-6-phosphate isomerase/mannose-1-phosphate guanyl transferase	40.9% ....
Cellular Processes			
Cell division			
Aq698	acrE	acriflavin resistance protein AcrE	24.8% ....
Aq1275	cafA	cytoplasmic axial filament protein	28.5% ....
Aq523	ftsA	cell division protein FtsA	31.9% ....
Aq936	ftsH	cell division protein FtsH	51.1% ....
Aq1139	ftsW	cell division protein FtsW	30.8% ...
Aq920	ftsY	cell division protein FtsY	35.2% ...
Aq525	ftsZ	cell division protein FtsZ	48.6% ....
Aq761	gidA1	glucose inhibited division protein A	50.2% ....
Aq691	gidA2	glucose inhibited division protein A	57.5% ....
Aq1582	gidB	glucose inhibited division protein B	39.4% ....
Aq1718	maf	MAF protein	44.9% ....
Aq1887	mesJ	cell cycle protein MesJ	27.7% ....
Aq878	minC	septum site-determining protein MinC	39.4% ..
Aq1217	minD1	septum site-determining protein MinD	33.1% ....
Aq877	minD2	septum site-determining protein MinD	54.5% ....
Aq845	mreB	rod shape determining protein MreB	57.4% ....
Aq025	rodA	rod shape determining protein RodA	37.6% ....
Aq1130	sufI	periplasmic cell division protein (SufI)	28.1% ....
Chaperones			
Aq154	ctaB	cytochrome c oxidase assembly factor	38.8% ....
Aq1735	dnaJ1	chaperone DnaJ	41.3% ....
Aq703	dnaJ2	chaperone DnaJ	45.1% ....
Aq996	dnaK	Hsp70 chaperone DnaK	59.1% ....
Aq433	grpE	heat shock protein GrpE	38.8% ....
Aq192	hslU	chaperone HslU	57.5% ....
Aq1283	hspC	small heat shock protein (class I)	31.0% ....
Aq1991	htpX	heat shock protein X	51.1% ....
Aq2200	mopA	GroEL	64.4% ....
Aq2199	mopB	GroES	56.2% ...
Detoxification			
Aq486	ahpC1	alkyl hydroperoxide reductase	49.2% ....
Aq858	ahpC2	alkyl hydroperoxide reductase	53.4% ....
Aq685	arsC	arsenate reductase	50.0% ....
Aq136	cpx	cytochrome c peroxidase	48.9% ....
Aq1005	cutA	periplasmic divalent cation tolerance protein	47.0% ....
Aq1499	sodA	superoxide dismutase (Fe/Mn family)	34.2% ....
Aq1050	sodC1	superoxide dismutase (Cu/Zn)	39.5% ....
Aq238	sodC2	superoxide dismutase (Cu/Zn)	39.2% ....
Aq488	tpx	thiol peroxidase	39.5% ....
Motility			
Aq833	flgA	flagellar protein FlgA	
Aq1184	flgB	flagellar basal body rod protein FlgB	
Aq1183	flgC	flagellar biosynthesis FlgC	39.4% ....
Aq1859	flgE	flagellar hook protein FlgE	30.8% ....
Aq2051	flgG1	flagellar hook basal-body protein FlgG	32.8% ....
Aq834	flgG2	flagellar hook basal-body protein FlgG	50.4% ....
Aq1714	flgH	flagellar L-ring protein FlgH	31.9% ....
Aq1713	flgI	flagellar P-ring protein FlgI	46.9% ....
Aq1662	flgK	flagellar hook associated protein FlgK	21.9% ....
Aq1663	flgL	flagellar hook associated protein FlgL	27.1% ..
Aq1212	flhA	flagellar export protein	44.0% ....
Aq2014	flhB	flagellar biosynthetic protein FlhB	39.8% ....
Aq1214	flhF	flagellar biosynthesis FlhF	28.7% ....
Aq1998	fliC	flagellin	59.4% ....

Aq2001	fliD	flagellar hook associated protein FliD	24.3% ..
Aq1182	fliF	Flagellar M-ring protein	32.0% ....
Aq653	fliG	flagellar switch protein FliG	35.9% ....
Aq1595	fliI	flagellar export protein	44.6% ....
Aq1860	fliL	flagellar biosynthesis FliL	30.6% ....
Aq1539	fliN	flagellar switch protein FliN	42.9% ...
Aq1920	fliP	flagellar biosynthetic protein FliP	47.7% ....
Aq1962	fliQ	flagellar biosynthesis protein FliQ	45.5% ....
Aq1961	fliR	flagellar biosynthetic protein FliR	29.7% ....
Aq2002	fliS	flagellar protein FliS	30.8% ....
Aq1003	motA	flagellar motor protein MotA	35.0% ....
Aq1002	motB1	flagellar motor protein MotB	36.8% ....
Aq1001	motB2	flagellar motor protein MotB-like	27.5% ....
Secretion			
Aq1720	ffh	signal recognition particle receptor protein	49.1% ....
Aq1288	gspD	general secretion pathway protein D	27.5% ....
Aq1474	gspE	general secretion pathway protein E	48.8% ....
Aq418	gspG	general secretion pathway protein G	50.7% ....
Aq955	lepB	type-I signal peptidase	33.9% ....
Aq1837	lsp	lipoprotein signal peptidase	37.4% ....
Aq1271	mpp	processing protease	28.7% ....
Aq747	pilC1	fimbrial assembly protein PilC	37.4% ....
Aq1285	pilC2	fimbrial assembly protein PilC	28.9% ....
Aq1601	pilD	type 4 prepilin peptidase	34.8% ....
Aq745	pilT	twitching motility protein PilT	51.4% ....
Aq2151	pilU	twitching mobility protein	41.6% ....
Aq1870	secA	preprotein translocase SecA subunit	44.9% ....
Aq973	secD	protein export membrane protein SecD	36.0% ....
Aq1602	secF	protein-export membrane protein	41.4% ....
Aq079	secY	preprotein translocase SecY	44.2% ....
Aq2080	sppA	proteinase IV	43.4% ....
Aq1971	tapB	type IV pilus assembly protein TapB	42.2% ....
Aq1340	tig	trigger factor	27.4% ....
Central Intermediary Metabolism			
One-carbon metabolism			
Aq1429	metF	5,10-methylenetetrahydrofolate reductase	43.3% ....
Aq1154	metK	S-adenosylmethionine synthetase	49.2% ....
Aq1180	sahH	S-adenosylhomocysteine hydrolase	60.9% ....
Cytoplasmic polysaccharides			
Aq1407	bcsA	cellulose synthase catalytic subunit	39.5% ....
Aq1401	celY	endoglucanase fragment	33.0% ..
Aq721	glgA	glycogen synthase	38.1% ....
Aq722	glgB	1,4-alpha-glucan branching enzyme	56.5% ....
Aq717	glgP	glycogen phosphorylase	37.0% ....
Aq723	malM	4-alpha-glucanotransferase (amylomaltase)	43.4% ....
Tri-carboxylic acid cycle			
Aq1784	aco	aconitase	36.1% ...
Aq1195	forA1	ferredoxin oxidoreductase alpha subunit	31.5% ...
Aq1167	forA2	ferredoxin oxidoreductase alpha subunit	32.3% ....
Aq1196	forB1	ferredoxin oxidoreductase beta subunit	29.6% ....
Aq1168	forB2	ferredoxin oxidoreductase beta subunit	31.5% ...
Aq1200	forG1	ferredoxin oxidoreductase gamma subunit	34.5% ....
Aq1169	forG2	ferredoxin oxidoreductase gamma subunit	34.5% ...
Aq594	frdA	fumarate reductase flavoprotein subunit	51.4% ....
Aq553	frdB1	reductase iron-sulfur subunit	35.2% ....
Aq655	frdB2	fumarate reductase iron-sulfur subunit	35.1% ....
Aq1780	fumB	fumarate hydratase (fumarase)	46.4% ....
Aq1679	fumX	C-terminal fumarate hydratase, class I	40.4% ....
Aq150	gltA	citrate synthase	33.0% ....
Aq1512	icd	isocitrate dehydrogenase	46.0% ....
Aq1782	mdh1	malate dehydrogenase	49.8% ....
Aq1665	mdh2	malate dehydrogenase	46.9% ...
Aq1614	oadA	oxaloacetate decarboxylase alpha chain	50.1% ....
Aq1306	sucC1	succinyl-CoA ligase beta subunit	35.1% ....
Aq1620	sucC2	succinyl-CoA ligase beta subunit	52.9% ....
Aq1888	sucD1	succinyl-CoA ligase alpha subunit	41.7% ...
Aq1622	sucD2	succinyl-CoA ligase alpha subunit	65.7% ....
Phosphate			
Aq1351	phoH	phosphate starvation-inducible protein	47.1% ....
Aq1547	ppa	inorganic pyrophosphatase	56.5% ....
Aq891	ppx	exopolyphosphatase	33.6% ....
Polyamines			
Aq728	speC	ornithine decarboxylase	30.9% ....
Aq062	speE	spermidine synthase	48.4% ....
Sulfur			
Aq1081	cysD	sulfate adenylyltransferase	46.7% ....
Aq1076	rhdA	thiosulfate sulfurtransferase	32.3% ....
Aq1799	rhdA	thiosulfate sulfurtransferase	31.7% ....
Aq455	sor	sulfur oxygenase reductase	36.7% ....
Aq1803	soxB	sulfur oxidation protein SoxB	41.3% ....
Cofactor Biosynthesis			
Lipoic acid biosynthesis			
Aq1355	lipA	Lipoic acid synthetase	48.9% ...
Biotin			
Aq170	bioA	DAPA aminotransferase	51.7% ....
Aq975	bioB	biotin synthetase	42.0% ....
Aq557	bioD	dethiobiotin synthetase	41.5% ....
Aq626	bioF	8-amino-7-oxononanoate synthase	45.1% ....
Aq1659	bioW	6-carboxyhexanoate-CoA ligase (pimeloyl CoA synthase)	47.3% ....
Aq566	birA	biotin [acetyl-CoA-carboxylase] ligase	37.5% ....
Folic acid			
Aq2045	folC	folylpolyglutamate synthetase	31.8% ....
Aq1898	folD	methylenetetrahydrofolate dehydrogenase	53.2% ....
Aq239	folE	GTP cyclohydrolase I	57.1% ....
Aq162	folK	folate biosynthesis 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase	43.7% ....
Aq1468	folP	dihydropteroate synthase	45.8% ....
Aq1144	pabB	p-aminobenzoate synthetase	41.5% ...
Aq1606	pabC	aminodeoxychorismate lyase	29.0% ....
Heme			
Aq207	cobA	uroporphyrin-III c-methyltransferase	52.1% ....
Aq1237	cysG	siroheme synthase	36.9% ....
Aq334	dcuP	uroporphyrinogen decarboxylase	41.4% ....
Aq816	gsa	glutamate-1-semialdehyde aminotransferase	56.5% ....
Aq1279	hemA	glutamyl tRNA reductase (delta-aminolevulinate synthase)	38.7% ....
Aq2109	hemB	porphobilinogen synthase	64.5% ....
Aq263	hemC	porphobilinogen deaminase	53.1% ....
Aq1424	hemF	oxygen-independent coproporphyrinogen III oxidase	33.1% ....
Aq2015	hemG	protoporphyrinogen oxidase	30.3% ...
Aq948	hemH	ferrochelatase	46.4% ....
Aq099	hemK	protoporphyrinogen oxidase	32.2% ....
Aq2124	hemN	oxygen-independent coproporphyrinogen II	50.2% ....
Molybdopterin			
Aq2183	moaA2	molybdenum cofactor biosynthesis protein A	47.0% ....
Aq527	moaC	molybdenum cofactor biosynthesis moaC	45.0% ..
Aq2181	moaE	molybdopterin converting factor subunit 2	39.3% ...
Aq1326	mobB	molybdopterin-guainine dinucleotide biosynthesis protein B	44.4% .
Aq030	moeA1	molybdenum cofactor biosynthesis protein A	36.8% ....
Aq1329	moeB	molybdopterin biosynthesis protein MoeB	54.1% ....
Aq061	mog	molybdenum cofactor biosynthesis MOG	55.5% ....
Aq049	phhB	pterin-4a-carbinolamine dehydratase	37.9% ....
Panthenate			
Aq815	dfp	pantothenate metabolism flavoprotein	41.2% ....
Aq1973	panB	3-methyl-2-oxobutanoate hydroxymethyltransferase	45.5% ....
Aq2132	panC	pantothenate synthetase	47.4% ....
Aq476	panD	aspartate 1-decarboxylase	46.0% ....
Pyridine nucleotides			
Aq1889	nadA	quinolinate synthetase A	44.3% ....
Aq777	nadB	L-aspartate oxidase	36.7% ...
Aq869	nadC	quinolinate phosphoribosyl transferase	47.0% ....
Aq959	nadE	NH(3)-dependent NAD+ synthetase	39.6% ....
Pyridoxal phosphate			
Aq852	pdxA	pyridoxal phosphate biosynthetic protein PdxA	36.8% ....
Aq1423	pdxJ	pyridoxal phosphate synthetase	88.2% ....
Quinones			
Aq895	ispB	octoprenyl-diphosphate synthase	35.7% ....
Aq052	ubiA	4-hydroxybenzoate octaprenyltransferase	41.4% ....
Riboflavin			
Aq350	ribA	GTP cyclohydrolase II	61.7% ....
Aq1707	ribC	riboflavin synthase alpha chain	45.3% ....
Aq138	ribD1	riboflavin specific deaminase	46.0% ....
Aq436	ribD2	riboflavin specific deaminase	42.9% ....
Aq139	ribF	riboflavin kinase	38.4% ....
Aq132	ribH	riboflavin synthase beta subunit	51.0% ....
Thiamine			
Aq1204	thiC	thiamine biosynthesis protein	67.1% ....
Aq1960	thiD	HMP-P kinase	40.5% ....
Aq1366	thiE1	thiamine phosphate synthase	36.3% ....
Aq558	thiE2	thiamine phosphate synthase	39.5% ....
Aq2178	thiG	thiamine biosynthesis, thiazole moiety	52.5% ....
Aq2119	thiL	thiamine monophosphate kinase	34.5% ....
Thio- and glutaredoxin			
Aq443	gua	glutaredoxin-like protein	33.8% ....
Aq1916	trxA1	thioredoxin	58.9% ...
Aq1811	trxA2	thioredoxin	32.2% ..
Aq500	trxB	thioredoxin reductase	39.8% ....
Energy Metabolism			
Aq1342	gph	phosphoglycolate phosphatase	33.9% ....
ATP-Proton Motive Force			
Aq679	atpA	ATP synthase F1 alpha subunit	64.3% ....
Aq179	atpB	ATP synthase F0 subunit a	36.4% ....
Aq673	atpC	ATP synthase F1 epsilon subunit	37.4% ....
Aq2038	atpD	ATP synthase F1 beta subunit	67.4% ....
Aq177	atpE	ATP synthase F0 subunit c	53.8% ...
Aq1586	atpF1	ATP synthase F0 subunit b	26.3% ....
Aq1587	atpF2	ATP synthase F0 subunit b	25.5% ....
Aq2041	atpG	ATP synthase F1 gamma subunit	39.9% ....
Aq1588	atpH	ATP synthase F1 delta chain	28.1% ....
Dehydrogenases			
Aq1362	adh1	alcohol dehydrogenase	35.4% ....
Aq1240	adh2	alcohol dehydrogenase	28.8% ....
Aq186	aldH1	aldehyde dehydrogenase	41.9% ....
Aq227	aldH2	aldehyde dehydrogenase	28.0% ....
Aq1145	dhaT	1,3 propanediol dehydrogenase	36.6% ....
Aq232	dhsU	flavocytochrome C sulfide dehydrogenase	33.6% ....
Aq1769	dld1	D-lactate dehydrogenase	45.3% ....
Aq1234	dmsA	DMSO reductase chain A	25.0% ....
Aq1232	dmsB	DMSO reductase chain B	38.4% ...
Aq1231	dmsC	DMSO reductase chain C	29.5% .
Aq1051	fdhE	formate dehydrogenase formation protein FdhE	25.9% ....
Aq1039	fdoG	formate dehydrogenase alpha subunit	50.0% ....
Aq1046	fdoH	formate dehydrogenase beta subunit	45.7% ....
Aq1049	fdoI	formate dehydrogenase gamma subunit	38.4% ....
Aq1903	gcsP1	glycine dehydrogenase (decarboxylating)	49.6% ....
Aq1109	gcsP2	glycine dehydrogenase (decarboxylating)	46.8% ....
Aq1639	glpC	oxido/reductase iron sulfur protein	27.1% ....
Aq395	hdrA	heterodisulfide reductase subunit A	39.7% ....
Aq400	hdrB	heterodisulfide reductase subunit B	32.5% ....
Aq398	hdrC	heterodisulfide reductase subunit C	35.7% .
Aq961	hdrD	heterodisulfide reductase	29.5% ....
Aq038	hibD	3-hydroxyisobutyrate dehydrogenase	34.6% ....
Aq727	ldhA	D-lactate dehydrogenase	33.5% ....
Aq736	lpdA	dihydrolipoamide dehydrogenase	37.0% ....
Aq217	narB	nitrate reductase narB	39.1% ....
Aq206	nirB	nitrite reductase (NAD(P)H) large subunit	35.3% ...
Aq835	nox	NADH oxidase	33.1% ....
Aq024	nsd	nucleotide sugar dehydrogenase	47.0% ....
Aq135	nueM	NADH dehydrogenase (ubiquinone)	28.2% ....
Aq1010	udh	dehydrogenase	29.7% ....
Electron transport			
Aq2191	coxA1	cytochrome c oxidase subunit I	42.4% ....
Aq2192	coxA2	cytochrome c oxidase subunit I	38.1% ....
Aq2190	coxB	cytochrome c oxidase subunit II	27.4% ....
Aq2188	coxC	cytochrome c oxidase subunit III	28.6% ....
Aq153	ctaA	heme O oxygenase	28.1% ....
Aq042	cyc	cytochrome c	25.8% ...
Aq792	cycB1	cytochrome c552	29.9% ..
Aq1550	cycB2	cytochrome C552	38.7% ....
Aq1357	cydA	cytochrome oxidase d subunit I	38.8% ....
Aq1358	cydB	cytochrome oxidase d subunit II	31.2% ....
Aq067	dmsB	dimethylsulfoxide reductase chain B	40.2% ....
Aq235	fccB'	sulfide dehydrogenase, flavoprotein subunit	38.0% ...
Aq919a	fdx1	ferredoxin	37.1% ...
Aq1171a	fdx2	ferredoxin	43.9% ..
Aq1192a	fdx3	ferredoxin	35.0% ...
Aq108a	fdx4	ferredoxin	56.6% ...
Aq211	fhp	flavohemoprotein	43.4% ....
Aq2096	floX	flavodoxin	32.5% ....
Aq045	petA	Rieske-I iron sulfur protein	34.3% ....
Aq044	petB	cytochrome b	38.3% ...
Aq234	soxF	Rieske-I iron sulfur protein	29.0% ....
Aq2186	sqr	sulfide-quinone reductase	41.0% ....
Glycolysis and gluconeogenesis			
Aq484	eno	enolase	65.0% ....
Aq1390	fba	fructose-1,6-bisphosphate aldolase class II	39.9% ....
Aq1065	gap	glyceraldehyde-3-phosphate dehydrogenase	59.5% ....
Aq434	glpK	glycerol kinase	51.0% ....
Aq1744	gpmA	phosphoglycerate mutase	27.9% ....
Aq1634	gpsA	glycerol-3-phosphate dehydrogenase (NAD+)	40.5% ....

Aq1708	pfkA	phosphofructokinase	49.4% ....
Aq750	pgi	glucose-6-phosphate isomerase	37.8% ....
Aq118	pgk	phosphoglycerate kinase	54.5% ....
Aq1990	pgmA	phosphoglycerate mutase	33.2% ....
Aq501	pmu	phosphoglucomutase/phosphomannomutase	37.8% ....
Aq2142	ppsA	phosphoenolpyruvate synthase	56.3% ....
Aq1520	pycA	pyruvate carboxylase c-terminal domain	46.6% ....
Aq1517	pycB	pyruvate carboxylase n-terminal domain	57.1% ....
Aq360	timA	triose phophate isomerase	52.2% ....
Hydrogenase			
Aq665	hoxZ	Ni/Fe hydrogenase B-type cytochrome subunit	40.4% ....
Aq667	hupD	HupD hydrogenase related function	40.9% ....
Aq666	hupE	HupE hydrogenase related function	38.3% ....
Aq1021	hypA	hydrogenase accessory protein HypA	39.8% ....
Aq671	hypB	hydrogenase expression/formation protein B	50.6% ....
Aq1157	hypD	hydrogenase expression/formation protein HypD	56.1% ....
Aq662	mbhL1	hydrogenase large subunit	50.6% ....
Aq960	mbhL2	hydrogenase large subunit	44.3% ....
Aq804	mbhL3	hydrogenase large subunit	27.9% ....
Aq660	mbhS1	hydrogenase small subunit	66.6% ....
Aq965	mbhS2	hydrogenase small subunit	51.3% ....
Aq802	mbhS3	hydrogenase small subunit	36.7% ....
Aq1591	shyS	soluble hydrogenase small subunit	41.6% ....
Sugar metabolism			
Aq968	cbbE2	ribulose-5-phosphate 3-epimerase	47.2% ....
Aq1658	fucA1	fuculose-1-phosphate aldolase	31.8% ....
Aq1979	fucA2	fuculose-1-phosphate aldolase	29.7% ....
Aq498	gnd	6-phosphogluconate dehydrogenase	45.2% ....
Aq497	gsdA	glucose-6-phosphate 1-dehydrogenase	32.3% ....
Aq1138	rpiB	ribose 5-phosphate isomerase B	54.5% ....
Aq119	talC	transaldolase	71.1% ....
Aq1765	tktA	transketolase	52.4% ....
NADH dehydrogenase			
Aq1385	nuoA1	NADH dehydrogenase I chain A	42.0% ....
Aq1310	nuoA2	NADH dehydrogenase I chain A	44.9% ....
Aq1312	nuoB	NADH dehydrogenase I chain B	60.1% ...
Aq551	nuoD1	NADH dehydrogenase I chain D	37.7% ....
Aq1314	nuoD2	NADH dehydrogenase I chain D	42.2% ....
Aq574	nuoE	NADH dehydrogenase I chain E	36.8% ....
Aq573	nuoF	NADH dehydrogenase I chain F	20.5% ..
Aq437	nuoG	NADH dehydrogenase I chain G	35.4% ..
Aq1315	nuoH1	NADH dehydrogenase I chain H	41.0% ....
Aq1373	nuoH2	NADH dehydrogenase I chain H	42.1% ....
Aq1374	nuoH3	NADH dehydrogenase I chain H	38.9% ....
Aq1317	nuoI1	NADH dehydrogenase I chain I	30.5% ...
Aq1375	nuoI2	NADH dehydrogenase I chain I	29.2% ...
Aq1318	nuoJ1	NADH dehydrogenase I chain J	35.4% ....
Aq1377	nuoJ2	NADH dehydrogenase I chain J	30.6% ....
Aq1319	nuoK1	NADH dehydrogenase I chain K	51.1% ....
Aq1378	nuoK2	NADH dehydrogenase I chain K	48.4% ....
Aq1320	nuoL1	NADH dehydrogenase I chain L	39.0% ....
Aq866	nuoL2	NADH dehydrogenase I chain L	30.2% ...
Aq1379	nuoL3	NADH dehydrogenase I chain L	43.1% ....
Aq1321	nuoM1	NADH dehydrogenase I chain M	43.6% ....
Aq1382	nuoM2	NADH dehydrogenase I chain M	36.9% ....
Aq1322	nuoN1	NADH dehydrogenase I chain N	34.1% ....
Aq1383	nuoN2	NADH dehydrogenase I chain N	32.8% ....
Lipid metabolism			
Aq2058	aas	2-acylglycerophosphoethanolamine acyltransferase	37.1% ....
Aq1206	accA	acetyl-CoA carboxylase alpha subunit	57.1% ...
Aq1363	accB	biotin carboxyl carrier protein	44.6% ....
Aq1664	accC1	biotin carboxylase	54.4% ....
Aq1470	accC2	biotin carboxylase	56.5% ....
Aq445	accD	acetyl-CoA carboxyltransferase beta subunit	56.9% ....
Aq1717a	acpP	acyl carrier protein	71.2% ....
Aq813	acpS	holo-[acyl-carrier protein] synthase	30.8% ....
Aq2104	acs	acetyl-coenzyme A synthetase	54.0% ....
Aq2103	acs'	acetyl-coenzyme A synthetase c-terminal fragment	61.2% ....
Aq1249	cds	phosphatidate cytidylyltransferase	29.2% ....
Aq1737	cfa	cyclopropane-fatty-acyl-phospholipid synthase	37.5% ....
Aq892	fabD	malonyl-CoA:Acyl carrier protein transacylase	42.1% ....
Aq1717	fabF	3-oxoacyl-[acyl-carrier-protein] synthase II	58.4% ....
Aq1716	fabG	3-oxoacyl-[acyl-carrier-protein] reductase	52.9% ....
Aq1099	fabH	3-oxoacyl-[acyl-carrier-protein] synthase III	47.0% ....
Aq1552	fabI	enoyl-[acyl-carrier-protein] reductase (NADH)	49.6% ....
Aq056	fabZ	(3R)-hydroxymyristoyl-(acyl carrier protein) dehydratase	58.7% ....
Aq999	fadD	long-chain-fatty-acid CoA ligase	30.0% ...
Aq1638	lplA	lipoate-protein ligase A	28.1% ..
Aq958	pgsA	phosphotidylglycerophosphate synthase	37.3% ..
Aq2154	pgsA	phosphotidylglycerophosphate synthase	38.9% ...
Aq1101	plsX	PlsX protein	43.7% ....
Purines, Pyrimidines, Nucleotides and Nucleosides			
Aq094	nrdA	ribonucleotide reductase alpha chain	35.0% ....
Aq1505	nrdF	ribonucleotide reductase beta chain	36.2% .
Purines			
Aq568	deoD	purine nucleoside phosphorylase	33.1% ....
Aq236	guaA	GMP synthase	58.4% ....
Aq2023	guaB	inosine monophosphate dehydrogenase	65.4% ....
Aq544	hpt	hypoxanthine-guanine phosphoribosyltransferase	48.2% ....
Aq078	kad	adenylate kinase	50.0% ....
Aq1590	ndk	nucleoside diphosphate kinase	48.2% ....
Aq1636	prs	phosphoribosylpyrophosphate synthetase	55.2% ....
Aq1290	purA	adenylosuccinate synthetase	49.2% ....
Aq597	purB	adenylosuccinate lyase	52.4% ....
Aq2117	purC	phosphoribosylaminoimidazole-succinocarboxamide synthase	52.5% ....
Aq742	purD	phosphoribosylamine-glycine ligase	54.2% ....
Aq1178	purE	phosphoribosylaminoimidazole carboxylase	64.6% ....
Aq1175	purF	amidophosphoribosyltransferase	42.7% ....
Aq1963	purH	phosphoribosylaminoimidazolecarboxamide formyltransferase	48.2% ....
Aq245	purK	phosphoribosyl aminoimidazole carboxylase	35.6% ....
Aq1836	purL	phosphoribosylformylglycinamidine synthase II	49.3% ...
Aq769	purM	phosphoribosylformylglycinamidine cyclo-ligase	50.0% ....
Aq857	purN	phosphoribosylglycinamide formyltransferase	48.3% ....
Aq1105	purQ	phosphoribosyl formylglycinamidine synthase I	51.1% ....
Aq1818	purU	formyltetrahydrofolate deformylase	56.3% ....
Pyrimidines			
Aq410	carA	carbamoyl phosphate synthetase small subunit	52.2% ....
Aq1172	carB	carbamoyl-phosphate synthase large subunit	60.7% ....
Aq2101	carB	carbamoyl-phosphate synthase, large subunit	63.1% ....
Aq2153	cmk	cytidylate kinase	38.5% ....
Aq1607	dcd	deoxycytidine triphosphate deaminase	39.5% ....
Aq220	dut	deoxyuridine 5'triphosphate nucleotidohydrolase	
Aq409	pyrB	aspartate carbamoyltransferase catalytic chain	42.0% ....
Aq806	pyrC	dihydroorotase	37.3% ...
Aq046	pyrD	dihydroorotase dehydrogenase	50.5% ....
Aq1305	pyrDB	dihydroorotate dehydrogenase electron transfer subunit	34.7% ....
Aq1580	pyrF	orotidine-5'-phosphate decarboxylase	37.2% ....
Aq1334	pyrG	CTP synthetase	57.5% ....
Aq713	pyrH	UMP kinase	62.1% ....
Aq640	thy	thymidylate synthase complementing protein	30.5% ...
Aq969	tmk	thymidylate kinase	35.1% ....
Aq1907	umpS	uridine 5-monophosphate synthase	42.1% ....
Aq2163	uraP	uracil phosphoribosyltransferase	42.0% ....
Regulation			
Aq1058	acrR1	transcriptional regulator (TetR/AcrR family)	34.1% ...
Aq2179	acrR2	transcriptional regulator (TetR/AcrR family)	31.0% ...
Aq281	acrR3	transcriptional regulator (TetR/AcrR family)	29.7% ....
Aq1387	arsR	transcriptional regulator (ArsR family)	35.3% ....
Aq1724	degT	transcriptional regulator (DegT/DnrJ/Eryc1 family)	34.1% ....
Aq534	draG	ADP-ribosylglycohydrolase	32.1% ....
Aq831	exsB	trans-regulatory protein ExsB	38.5% ....
Aq490	fnr	transcriptional regulator (Crp/Fnr family)	29.5% ....
Aq1207	furR1	transcriptional regulator (FurR family)	37.9% ....
Aq1418	furR2	transcriptional regulator (FurR family)	34.6% ....
Aq213	glnBi	PII-like protein GlnBi	48.0% ..
Aq1908	hflX	GTP-binding protein HflX	40.3% ....
Aq1115	hksP1	histidine kinase sensor protein	27.7% ..
Aq316	hksP2	histidine kinase sensor protein	28.1% ....
Aq905	hksP3	histidine kinase sensor protein	23.6% ...
Aq231	hksP4	histidine kinase sensor protein	28.2% ....
Aq1156	hoxX	hydrogenase regulation HoxX	46.7% ....
Aq093	hth	transcriptional regulator (H-T-H)	50.2% ....
Aq1019	hypE	hydrogenase expression/formation protein	44.3% ....
Aq672	hypF	transcriptional regulatory protein HypF	44.8% ....
Aq764	iclR	transcriptional regulator (IclR family)	30.4% ....
Aq638	lysR1	transcriptional regulator (LysR family)	32.8% ....
Aq1038	lysR2	transcriptional regulator (LysR family)	28.9% ....
Aq702	merR	transcriptional regulator (MerR family)	32.8% ....
Aq218	nifA	transcriptional regulator (NifA family)	42.8% ....
Aq1117	ntrC1	transcriptional regulator (NtrC family)	41.0% ....
Aq1792	ntrC2	transcriptional regulator (NtrC family)	40.2% ....
Aq230	ntrC3	transcriptional regulator (NtrC family)	40.0% ....
Aq164	ntrC4	transcriptional regulator (NtrC family)	38.3% ....
Aq2069	obg	GTP-binding protein	54.9% ..
Aq319	phoB	transcriptional regulator (PhoB-like)	41.6% ....
Aq906	phoU	transcriptional regulator (PhoU-like)	41.9% ....
Aq844	spoT	(p)ppGpp 3-pyrophosphohydrolase	47.2% ...
Aq1496	xylR	transcriptional regulator (NagC/XylR family)	29.3% ....
DNA Replication and Repair			
Aq358	dinG	ATP-dependent helicase (DinG family)	27.9% ....
Aq322	dnaA	chromosome replication initiator protein DnaA	36.5% ....
Aq1472	dnaB	replicative DNA helicase	40.3% ....
Aq910	dnaC	DNA replication protein DnaC	26.4% ....
Aq1008	dnaE	DNA polymerase III alpha subunit	41.9% ...
Aq1493	dnaG	DNA primase	39.8% ...
Aq1882	dnaN	DNA polymerase III beta chain	32.1% ....
Aq932	dnaQ	DNA polymerase III epsilon subunit	40.0% ....
Aq1855	dnaX	DNA polymerase III gamma subunit	36.6% ....
Aq1422	dpbF	DNA polymerase beta family	39.1% ....
Aq1693	dplF	N-terminus of phage SPO1 DNA polymerase	37.3% ....
Aq980	gyrA	DNA gyrase A subunit	43.6% ....
Aq1026	gyrB	gyrase B	55.2% ..
Aq2057	helX	DNA helicase	49.7% ....
Aq1484a	himA	DNA binding protein HU	40.2% ....
Aq2174	ihfB	integration host factor beta subunit	35.8% ....
Aq1394	lig	DNA ligase (ATP dependent)	50.8% ....
Aq633	ligA	DNA ligase (NAD dependent)	45.7% ....
Aq1578	mutL	DNA mismatch repair protein MutL	72.3% ....
Aq308	mutS1	DNA mismatch repair protein MutS	77.5% ....
Aq1242	mutS2	DNA mismatch repair protein MutS	37.0% ....
Aq1449	mutT	8-OXO-dGTPase domain (mutT domain)	46.3% ....
Aq282	mutY1	endonuclease III	53.6% ....
Aq172	mutY2	endonuclease III	51.8% ....
Aq496	mutY3	endonuclease III	43.4% ....
Aq1629	nfo	deoxyribonuclease IV	39.0% ....
Aq710	nucI	thermococcal nuclease homolog	36.4% ....
Aq1495	ogt	O-6-methylguanine-DNA-alkyltransferase	36.9% ...
Aq1628	pol	DNA polymerase I 3'-5' exo domain	43.2% ....
Aq1967	polA	DNA polymerase I (PolI)	30.5% ....
Aq1610	radC	DNA repair protein RadC	39.0% ....
Aq2150	recA	recombination protein RecA	88.5% ....
Aq2053	recG	ATP-dependent DNA helicase RecG	38.9% ....
Aq2155	recJ	single-strand-DNA-specific exonuclease RecJ	31.8% ....
Aq561	recN	recombination protein RecN	27.7% ....
Aq1478	recR	recombination protein RecR	38.3% ....
Aq793	rep	ATP-dependent DNA helicase REP	33.4% ....
Aq1886	sbcD	ATP-dependent dsDNA exonuclease	29.9% ....
Aq064	ssb	single stranded DNA-binding protein	39.4% ...
Aq657	topA	topoisomerase I	39.6% ....
Aq1159	topG1	reverse gyrase	41.6% ....
Aq886	topG2	reverse gyrase	35.1% ....
Aq686	uvrA	repair excision nuclease subunit A	61.0% ....
Aq1856	uvrB	repair excision nuclease subunit B	53.9% ....
Aq2126	uvrC	repair excision nuclease subunit C	32.5% ....
Transcription			
RNA polymerase and transcription factors			
Aq613	deaD	ATP-dependent RNA helicase DeaD	42.3% ....
Aq357a	flgM	anti sigma factor FlgM	20.6% ....
Aq1218	fliA	RNA polymerase sigma factorFliA	37.2% ....
Aq259	nusA	transcription termination NusA	45.4% ...
Aq133	nusB	transcription termination NusB	32.3% ....
Aq1931	nusG	transcription antitermination protein NusG	46.3% ..
Aq873	rho	transcriptional terminator Rho	59.6% ....
Aq070	rpoA	RNA polymerase alpha subunit	40.4% ....
Aq1939	rpoB	RNA polymerase beta subunit	44.0% ....
Aq1945	rpoC	RNA polymerase beta prime subunit	46.9% ....
Aq1490	rpoD	RNA polymerase sigma factor RpoD	41.6% ....
Aq599	rpoN	RNA polymerase sigma factor RpoN	30.6% ....
Aq1452	rpoS	RNA polymerase sigma factor RpoS	40.5% ....
RNA modification			
Aq1816	ksgA	dimethyladenosine transferase	36.1% ....
Aq1067	miaA	tRNA delta-2-isopentenylpyrophosphate (IPP) transferase	38.2% ....
Aq411	pcnB1	poly A polymerase	28.5% ....
Aq2158	pcnB2	poly A polymerase	33.9% ..
Aq221	phpA	polyribonucleotide nucleotidyltransferase	45.0% ....
Aq894	queA	queuosine biosynthesis protein	46.9% ....
Aq946	rnc	RNase III	35.8% ....
Aq1955	rnhB	RNase HII	48.4% ....
Aq924	rnpH	RNase PH	64.0% ....
Aq1661	spoU	rRNA methylase SpoU	44.0% ....
Aq1308	tgt	queuine tRNA-ribosyltransferase	52.6% ....
Aq841	trm1	N2,N2-dimethylguanosine tRNA	

		methyltransferase	34.6% ....
Aq1489	trmD	tRNA guanine-N1 methyltransferase	42.9% ....
Aq749	truA	pseudouridine synthase I	33.1% ....
Aq705	truB	tRNA pseudouridine 55 synthase	38.2% ....
Aq1890	tsnR	rRNA methylase	36.4% ....
Aq2046	vacB	VacB protein (ribonuclease II family)	37.9% ....
Aq257	ygcA	RNA methyltransferase (TrmA-family)	28.8% ....
Translation			
Aq2131	fmt	methionyl-tRNA formyltransferase	45.7% ....
Aq247	gatA	glutamyl-tRNA(Gln) amidotransferase subunit A	53.6% ....
Aq461	gatB	glutamyl-tRNA(Gln) amidotransferase subunit B	48.8% ....
Aq2147a	gatC	glutamyl-tRNA(Gln) amidotransferase subunit C	41.1% ....
Aq346	pth	peptidyl-tRNA hydrolase	48.8% ....
Aminoacyl tRNA synthetases			
Aq1293	alaS	alanyl-tRNA synthetase	46.6% ....
Aq923	argS	arginyl-tRNA synthetase	39.4% ....
Aq1677	aspS	aspartyl-tRNA synthetase	51.3% ....
Aq1068	cysS	cysteinyl-tRNA synthetase	45.0% ....
Aq763	genX	lysyl-tRNA synthetase (genX) homolog	38.6% ....
Aq1221	gltX	glutamyl-tRNA synthetase	48.5% ....
Aq945	glyQ	glycyl-tRNA synthetase alpha subunit	61.9% ....
Aq2141	glyS	glycyl-tRNA synthetase beta subunit	37.1% ....
Aq122	hisS1	histidyl-tRNA synthetase	43.3% ....
Aq1155	hisS2	histidyl-tRNA synthetase	34.9% ....
Aq305	ileS	isoleucyl-tRNA synthetase	82.1% ..
Aq351	leuS	leucyl-tRNA synthetase alpha subunit	50.7% ....
Aq1770	leuS'	leucyl-tRNA synthetase beta subunit	47.2% ....
Aq1202	lysU	lysyl-tRNA synthetase	53.2% ...
Aq1257	metG	methionyl-tRNA synthetase alpha subunit	45.0% ....
Aq422	metG'	methionyl-tRNA synthetase beta subunit	64.2% ....
Aq953	pheS	phenylalanyl-tRNA synthetase alpha subunit	51.9% ....
Aq1730	pheT	phenylalanyl-tRNA synthetase beta subunit	35.4% ....
Aq365	proS	proline-tRNA synthetase	44.1% ....
Aq298	serS	seryl-tRNA synthetase	59.4% ....
Aq1667	thrS	threonyl-tRNA synthetase	48.5% ....
Aq992	trpS	tryptophanyl-tRNA synthetase	38.4% ....
Aq1751	tyrS	tyrosyl tRNA synthetase	56.2% ....
Aq1413	valS	valyl-tRNA synthetase	33.2% ....
Ribosomal Proteins			
Aq1935	rplA	ribosomal protein L01	57.9% ....
Aq013	rplB	ribosomal protein L02	46.9% ....
Aq009	rplC	ribosomal protein L03	53.8% ....
Aq011	rplD	ribosomal protein L04	51.3% ....
Aq1652	rplE	ribosomal protein L05	67.0% ....
Aq1649	rplF	ribosomal protein L06	46.2% ....
Aq2042	rplI	ribosomal protein L09	35.6% ....
Aq1936	rplJ	ribosomal protein L10	36.5% ....
Aq1933	rplK	ribosomal protein L11	71.4% ....
Aq1937	rplL	ribosomal protein L7/L12	75.4% ....
Aq1877	rplM	ribosomal protein L13	60.6% ....
Aq1654	rplN	ribosomal protein L14	59.5% ....
Aq1642	rplO	ribosomal protein L15	57.4% ....
Aq018	rplP	ribosomal protein L16	59.3% ....
Aq069	rplQ	ribosomal protein L17	48.7% ....
Aq1648	rplR	ribosomal protein L18	62.7% ....
Aq1954	rplS	ribosomal protein L19	59.8% ...
Aq952	rplT	ribosomal protein L20	63.5% ....
Aq016a	rplV	ribosomal protein L22	47.3% ....
Aq012	rplW	ribosomal protein L23	52.2% ....
Aq1653	rplX	ribosomal protein L24	50.8% ....
Aq1644	rpmD	ribosomal protein L30	46.4% ..
Aq1930a	rpmG	ribosomal protein L33	67.9% ....
Aq792a	rpmI	ribosomal protein L35	48.3% ....
Aq1485	rpsA	ribosomal protein S01	32.6% ....
Aq2007	rpsB	ribosomal protein S02	60.3% ....
Aq017	rpsC	ribosomal protein S03	54.0% ....
Aq072	rpsD	ribosomal protein S04	51.9% ....
Aq1645	rpsE	ribosomal protein S05	60.6% ...
Aq063	rpsF	ribosomal protein S06	32.7% ....
Aq1832	rpsG1	ribosomal protein S07	52.5% ....
Aq734	rpsG2	ribosomal protein S07	51.9% ....
Aq1651	rpsH	ribosomal protein S08	39.9% ....
Aq1878	rpsI	ribosomal protein S09	50.5% ...
Aq008	rpsJ	ribosomal protein S10	55.9% ....
Aq073	rpsK	ribosomal protein S11	60.7% ....
Aq735	rpsL1	ribosomal protein S12	78.9% ....
Aq1834	rpsL2	ribosomal protein S12	78.9% ....
Aq074	rpsM	ribosomal protein S13	61.9% ....
Aq1651a	rpsN	ribosomal protein S14	51.6% ....
Aq226a	rpsO	ribosomal protein S15	61.6% ....
Aq123	rpsP	ribosomal protein S16	36.6% ...
Aq020	rpsQ	ribosomal protein S17	59.6% ....
Aq064a	rpsR	ribosomal protein S18	48.5% ....
Aq015	rpsS	ribosomal protein S19	63.1% ..
Aq1767	rpsT	ribosomal protein S20	40.0% ...
Aq867a	rpsU	ribosomal protein S21	38.2% ....
Translation factors			
Aq1364	efp	elongation factor P	48.6% ....
Aq2114	eif	initiation factor eIF-2B alpha subunit	58.4% ...
Aq712	frr	ribosome recycling factor	43.0% ....
Aq001	fusA	elongation factor EF-G	91.9% ....
Aq075a	infA	initiation factor IF-1	69.1% ....
Aq2032	infB	initiation factor IF-2	48.5% ....
Aq1777	infC	initiation factor IF-3	53.6% ....
Aq876	prfA	peptide chain release factor RF-1	54.8% ....
Aq1840	prfB	peptide chain release factor RF-2	49.9% ....
Aq1033	selB	elongation factor SelB	30.4% ....
Aq715	tsf	elongation factor EF-Ts	35.8% ....
Aq005	tufA1	elongation factor EF-Tu	74.4% ....
Aq1928	tufA2	elongation Factor EF-Tu	73.9% ....
Protein modification			
Aq731	ccdA	cytochrome c-type biogenesis protein	32.0% ....
Aq579	def	polypeptide deformylase	41.4% ....
Aq2093	dsbC	thiol:disulfide interchange protein	27.6% ....
Aq055	hemX1	cytochrome c biogenesis protein	26.2% ....
Aq2043	hemX2	cytochrome c biogenesis protein	36.2% ....
Aq1053	nifS1	FeS cluster formation protein NifS	38.5% ....
Aq739	nifS2	FeS cluster formation protein NifS	45.5% ....
Aq1871	pmbA	peptide maturation	25.6% ....
Aq2102	prmA	ribosomal protein L11 methyltransferase	35.1% ....
Aq567	rimI	ribosomal-protein-alanine acetyltransferase	37.9% ....
Aq576	stpK	ser/thr protein kinase	30.8% ....
Aq152	tlpA	thiol disulfide interchange protein	37.6% ...
Proteases			
Aq1950	aprV	serine protease	26.5% ...
Aq1672	clpB	ATPase subunit of ATP-dependent protease	46.8% ....
Aq1296	clpC	ATP-dependent Clp protease	54.9% ....
Aq1339	clpP	ATP-dependent Clp protease proteolytic subunit	65.4% ....
Aq1337	clpX	ATP-dependent protease ATPase subunit clpX	66.1% ....
Aq1015	col	collagenase	41.3% ....
Aq1671	hslV	heat shock protein HsLV	57.6% ....
Aq1450	htrA	periplasmic serine protease	38.3% ....
Aq242	lon	Lon protease	50.6% ....
Aq076	map	methionyl aminopeptidase	44.1% ....
Aq1459	npr	neutral protease	27.7% ..
Aq2099	pepA	leucine aminopeptidase	39.5% ....
Aq1535	pepQ	xaa-pro dipeptidase	31.9% ....
Aq618	pfpI	protease I	41.8% ....
Aq797	prc	carboxyl-terminal protease	41.8% ....
Aq552	sms	ATP-dependent protease sms	46.2% ....
Aq2204	ymxG	processing protease	28.3% ....
Transport			
Aq1222	abcT1	ABC transporter	34.7% ....
Aq620	abcT2	ABC transporter	36.8% ....
Aq1095	abcT3	ABC transporter (ABC-2 subfamily)	34.4% ....
Aq1094	abcT4	ABC transporter	37.7% ....
Aq1097	abcT5	ABC transporter (hlyB subfamily)	45.5% ....
Aq417	abcT6	ABC transporter	51.8% ....
Aq413	abcT7	ABC transporter	51.5% ....
Aq297	abcT8	ABC transporter	49.3% ....
Aq2160	abcT9	ABC transporter	45.3% ....
Aq1531	abcT10	ABC transporter	36.4% ....
Aq2122	abcT11	ABC transporter	42.5% ....
Aq2137	abcT12	ABC transporter	38.2% ....
Aq1563	abcT13	ABC transporter (MsbA subfamily)	30.5% ....
Aq695	acrD1	cation efflux system (AcrB/AcrD/AcrF family)	22.7% ....
Aq1122	acrD2	cation efflux system (AcrB/AcrD/AcrF family)	32.0% ....
Aq469	acrD3	cation efflux system (AcrB/AcrD/AcrF family)	34.2% ....
Aq786	acrD4	cation efflux (AcrB/AcrD/AcrF family)	27.7% ....
Aq112	amtB	ammonium transporter	49.0% ....
Aq682	arsA1	anion transporting ATPase	41.5% ....
Aq343	arsA2	anion transporting ATPase	33.9% ....
Aq851	corA	Mg(2+) and Co(2+) transport protein	31.1% ....
Aq724	ctrA1	cation transporting ATPase (E1-E2 family)	30.7% ....
Aq1445	ctrA2	cation transporting ATPase (E1-E2 family)	28.1% ....
Aq1125	ctrA3	cation transporting ATPase (E1-E2 family)	43.8% ....
Aq1132	czcB1	cation efflux system (czcB-like)	23.7% ....
Aq1331	czcB2	cation efflux system (czcB-like)	26.9% ....
Aq468	czcB2	cation efflux system (czcB-like)	28.5% ....
Aq1073	czcD	cation efflux system (CzcD-like)	43.4% ....
Aq911	ebs	erythrocyte band 7 homolog	50.2% ....
Aq1062	emrB	major facilitator family transporter	28.3% ....
Aq1255	feoB	ferrous iron transport protein B	32.6% ....
Aq1330	gltP	proton/sodium-glutamate symport protein	35.6% ....
Aq1268	hvsT	high affinity sulfate transporter	29.4% ....
Aq1863	kch	potassium channel protein	30.1% ...
Aq1725	lepA	G-protein LepA	59.8% ....
Aq1229	mfiT	transporter (major facilitator family)	37.2% ....
Aq447	mgtC	Mg(2+) transport ATPase	36.2% ....
Aq1609	modA	molybdate periplasmic binding protein	38.2% ....
Aq086	modC	Molybdenum transport system permease	44.8% ....
Aq415	napA1	Na(+)/H(+) antiporter	27.6% ....
Aq929	napA2	Na(+)/H(+) antiporter	32.7% ....
Aq2030	napA3	Na(+)/H(+) antiporter	26.8% ....
Aq215	nasA	nitrate transporter	35.8% ...
Aq1441	oppA	transporter (extracellular solute binding protein family 5)	37.0% ....
Aq481	oppB	transporter (OppBC family)	46.2% ....
Aq1509	oppC	oligopeptide transport system permease	46.2% ....
Aq2019	pstA	phosphate transport system permease PstA	43.5% ....
Aq1055	pstB	phosphate transport ATP binding protein	68.1% ....
Aq2018	pstC	phosphate transport system permease protein C	45.2% ....
Aq2016	pstS	phosphate-binding periplasmic protein	52.4% ....
Aq2129	sbf	Na(+) dependent transporter (Sbf family)	34.9% ....
Aq098	secG	protein export membrane protein SecG	35.7% ....
Aq2077	snf	Na(+):neurotransmitter symporter (Snf family)	25.7% ....
Aq2106	ssf	Na(+):solute symporter (Ssf family)	47.4% .
Aq1988	tolQ	TolQ homolog	32.5% .
Aq1504	trk1	K+ transport protein homolog	40.6% ....
Aq031	trnS	transporter (Pho87 family)	46.8% ....
Uncategorized			
Aq1023	acuC1	acetoin utilization protein	36.9% ....
Aq2110	acuC2	acetoin utilization protein	38.6% ....
Aq158	apfA	AP4A hydrolase	36.6% ....
Aq458	bcp	bacterioferritin comigratory protein	40.6% ....
Aq542	bcpC	phosphonopyruvate decarboxylase	37.4% ....
Aq147	cobW	cobalamin synthesis related protein CobW	29.5% ...
Aq1303a	cspC	cold shock protein	67.2% ....
Aq1265	cstA	carbon starvation protein A	33.0% ....
Aq348	ctc	general stress protein Ctc	34.7% ....
Aq212	cynS	cyanate hydrolase	39.5% ....
Aq337	cysQ	CysQ protein	47.4% ....
Aq528	dedF	phenylacrylic acid decarboxylase	52.4% ....
Aq148	deoC	deoxyribose-phosphate aldolase	46.6% ....
Aq2095	dksA	dnaK suppressor protein	35.1% ....
Aq1994	era1	GTP-binding protein Era	49.7% ....
Aq1919	era2	GTP binding protein Era	43.0% ....
Aq1540	gcpE	GcpE protein	50.1% ....
Aq1052	gcsH1	glycine cleavage system protein H	28.6% ....
Aq1657	gcsH2	glycine cleavage system protein H	39.8% ...
Aq944	gcsH3	glycine cleavage system protein H	36.7% ...
Aq1108	gcsH4	glycine cleavage system protein H	44.5% ...
Aq1458	gcvT	aminomethyltransferase (glycine cleavage system T protein)	42.2% ....
Aq108b	hfq	host factor I	53.5% ....
Aq101	hly	hemolysin	33.7% ...
Aq2120	hlyC	hemolysin homolog protein	29.3% ....
Aq1091	hylA	hemolysin	33.5% ...
Aq708	hyuA	N-methylhydantoinase A	39.8% ....
Aq1925	hyuB	N-methylhydantoinase B	43.1% ....
Aq1579	iagB	invasion protein IagB	38.3% ...
Aq1983	imp2	myo-inositol-1(or 4)-monophosphatase	36.0% ....
Aq748	ispA	geranylgeranyl pyrophosphate synthase	40.7% ....
Aq1739	lytB	LytB protein	43.9% ....
Aq1977	masA	enolase-phosphatase E-1	42.3% ....
Aq1560	mglA1	gliding motility protein	42.4% ....
Aq1823	mglA2	gliding motility protein MglA	34.1% ....
Aq1789	mviB	'virulence factor' homolog MviB	29.7% ....
Aq587	neaC	N-ethylammeline chlorohydrolase	42.8% ....
Aq1820	nfeD	nodulation competitiveness protein NfeD	37.9% ....
Aq896	nifU	NifU protein	48.3% ....
Aq1300	omp	outer membrane protein	25.5% ....
Aq1507	omt	O-methyltransferase	39.5% ....
Aq967	ostA	organic solvent tolerance protein	22.0% ....
Aq141	pkcI	protein kinase C inhibitor (HIT family)	59.0% ....
Aq994	pncA	pyrazinamidase/nicotinamidase	39.1% ....
Aq057	sfsA	sugar fermentation stimulation protein	27.3% ....
Aq287	smb	small protein B	52.0% ....
Aq832	surE	stationary phase survival protein SurE	44.1% ....
Aq871	thdF	thiophene and furan oxidation protein	45.4% ....
Aq2021	tldD	TldD protein	40.[illegible] ....
Aq773	tly	hemolysin	41.[illegible] ....
[illegible]		[illegible]	[illegible] ....

from CheA are transferred to CheY, which then binds to the flagellar switch, altering the direction of flagellar rotation. Homologous chemotaxis systems are present in the archaea *Halobacterium salinarum*[29] and *Pyrococcus* sp. OT3 (H. Sizuya, personal communication), although the bacterial and archaeal flagellar apparatuses are not homologous[30]. The *M. jannaschii* genome also lacks homologues of known genes required for chemotaxis. Thus, either motility in *A. aeolicus* and *M. jannaschii* is undirected or input for controlling taxis is mediated through another, unidentified system. The most studied chemotaxis systems respond to sugars and amino acids, although responses to other inputs (for example, metals, redox potential, and light) may also occur. In contrast to all the organisms known to possess the classical chemotactic signal-transduction pathways, both *A. aeolicus* and *M. jannaschii* are obligate chemoautotrophs. Chemoautotrophs may respond to a different set of factors, such as concentrations of dissolved gas (CO_2, H_2 or O_2) or another critical parameter such as temperature.

In *E. coli*, the flagellar switch is essential for flagellar structure and function and coupling of chemotaxis signals. But the *A. aeolicus* genome encodes homologues of only two of the three *E. coli* proteins that make up the switch, FliG and FliN. Biochemical[31] and genetic[32] studies implicate the missing FliM protein as the receptor for phosphorylated CheY, the switch signal. The absence of both FliM and CheY in *A. aeolicus* supports the identification of FliM as the receptor for phosphorylated CheY in *E. coli*. This result also argues against a direct role for FliM in torque generation.

DNA replication and repair

The *A. aeolicus* primary replicative DNA polymerase, corresponding to the DNA polymerase III holoenzyme in *E. coli*, probably consists

Box 1 *Aquifex aeolicus* genome features

General

Length: 1,551,335 bp
G + C content: 43.4%
Protein-coding regions: 93%
Stable RNA: 0.8%
Non-coding repeats: (none significant)
Intergenic sequences: 6.2%

RNA

Ribosomal RNA	Chromosome coordinates
16S-23S-5S	672785-667770
16S-23S-5S	1192069-1197084

Transfer RNA
44 species (7 clusters, 28 single genes)

Other RNAs	Chromosome coordinates
tmRNA	1153844-1153498

Chromosomal coding sequences

849 similar to protein of known function (average length 1,066 bp)
256 similar to protein of unknown function (average length 898 bp)
407 unknown coding regions (average length 762 bp)
1,512 total (average length 956)

Extrachromosomal element (ECE)

Length: 39,456 bp
G + C content: 36.4%
Protein-coding regions: 53.5%

ECE-coding sequences

1 similar to proteins of known function (length 948 bp)
4 similar to proteins of unknown function (average length 667 bp)
27 unknown coding regions (average length 648 bp)

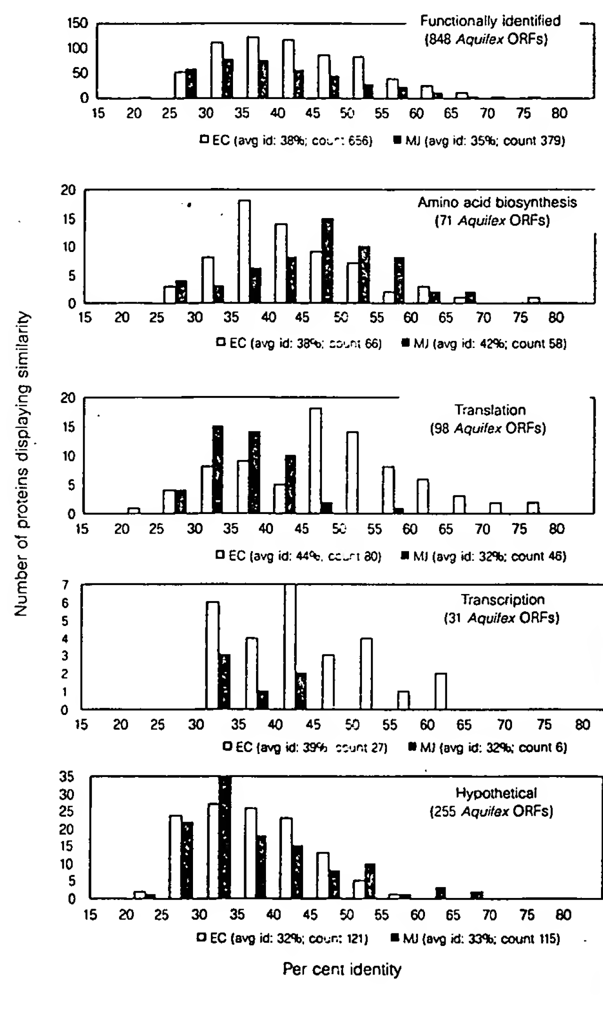


Figure 2 Histogram representation of the similarity of selected classes of predicted proteins to predicted proteins from the *E. coli* (EC) and *M. jannaschii* (MJ) genomes. Predicted *A. aeolicus* proteins representing each category were independently compared to sets of all potential polypeptides (≥100 amino acids) from the two genomes using FASTA[44]. If the top scoring alignment covered ≥80% of the length of the *A. aeolicus* protein, the score was plotted. There were more positives found in the *E. coli* genome in nearly every category. Hypothetical proteins (those identified by database match but of unknown function) are very similarly represented by *M. jannaschii* and *E. coli*. There are a small number of very highly conserved hypotheticals that are shared between *A. aeolicus* and *M. jannaschii*. Generally, biosynthetic categories show less discrimination than information-processing categories, which are clearly more *E. coli*-like. The variation in the apparent rates of evolution in different categories suggests that different phylogenies may be inferred depending on the sequence analysed. Within each graph, correspondence to *E. coli* is shown in white and *M. jannaschii* is shown in black. Avg id, average identity; count, number of proteins analysed.

of a core structure containing α- and ϵ-subunits, a γ-τ-subunit and an additional member of the γ-τ/δ'-family. A gene encoding a protein homologous to the β-sliding clamp was also found. This minimalistic complex lacks homologous θ-, δ-, χ- and ψ-subunits, as does the *Mycoplasma genitalium* holoenzyme[3]. Translation of the 54K (relative molecular mass) γ-τ-ATPase subunit may proceed without a programmed frameshift to produce a protein similar to the N-terminal region of the *E. coli* γ-subunit. DNA polymerase I is present as separate Klenow fragment and $5' \rightarrow 3'$ exonuclease subunits, encoded by two non-adjacent ORFs. Although the repair polymerase, DNA polymerase II, has not been found in *A. aeolicus*, one ORF (Aq1422) encodes a protein similar to the eukaryotic DNA repair polymerase-β. A member of the same family has been identified in *Thermus aquaticus*[33] and *Bacillus subtilis*.

Transcriptional and translational apparatuses

The transcriptional apparatus of *A. aeolicus* is similar to that of *E. coli* and lacks any components specific to the Eukarya or Archaea (Fig. 2). In addition to the core RNA polymerase α-, β-, and β'-subunits, four σ-factors which determine promoter specificity are present (Table 1). Several different families of bacterial transcriptional regulators were also identified, including two-component systems. All of the ribosomal proteins and elongation factors common to other bacteria are present, indicating that all bacteria-specific ribosomal proteins were present in the common ancestor of *Aquifex* and other bacteria. Also present are the four *sel* genes required for the cotranslational incorporation of selenocysteine. These latter genes are clustered in a 15-kilobase-pair segment that also encodes the biosynthetic and structural proteins for formate dehydrogenase, the only selenocysteine-containing protein identified. The gene that encodes selenocysteine transfer RNA, *selC*, is apparently cotranscribed with the genes encoding the formate dehydrogenase structural proteins.

A. aeolicus lacks glutaminyl-tRNA and asparaginyl-tRNA synthetases. The genes required for transamidation of glutamyl-$tRNA^{Gln}$ are present[34]. Charging of asparaginyl-tRNA is likely to proceed through the analogous reaction, as shown in halobacteria[35], although the genes(s) for that transamidase are unknown. The canonical methionyl- and leucyl-tRNA synthetases have only been seen previously as single polypeptide enzymes; however, in *A. aeolicus* the homologues appear fragmented into two subunits. In both cases, the genes that encode the N- and C-terminal portions are widely separated on the chromosome. No complete three-dimensional structural data are available for either methionyl- or leucyl-aminoacyl tRNA synthetases, but the subunit organization in the *A. aeolicus* aminoacyl-tRNA synthetases may reflect domain organization in the homologous proteins.

Thermophily

The *A. aeolicus* genome is the second completely sequenced genome of a hyperthermophile. By comparing the *A. aeolicus* and *M. jannaschii* genomes and contrasting them with the complete genomes of mesophiles, we can discover whether there are aspects of the genome or the encoded information that are diagnostic of hyperthermophiles. The G + C content of the stable RNAs is clearly indicative of the high growth temperature of the organism. This property can be used to identify stable RNAs against the relatively low G + C background of the *A. aeolicus* genome. The gene encoding tmRNA (or 10Sa RNA)[36], an RNA involved in tagging polypeptides translated from incomplete messenger RNAs for degradation, was located in this way.

Two genes for reverse gyrase are present in the genome. This is the only protein known to be present only in thermophiles. Other proteins, currently described as hypotheticals, may be diagnostic of hyperthermophiles but the data sets are not yet large enough to decide this with confidence.

Although features of stabilization may not be apparent in any given protein[37], a large enough data set may reveal general trends in amino-acid usage that are informative. Particularly important in this regard is inclusion of multiple genomes of hyperthermophiles so as not to allow the idiosyncracies of a single organism to bias the conclusions. As shown in Table 2, comparison of the amino-acid composition encoded by six genomes shows that use of individual amino acids can vary significantly from genome to genome. The data suggest trends that may be correlated with the thermostability of the encoded proteins. One apparent trend is that the hyperthemophile genomes encode higher levels of charged amino acids on average than mesophile genomes[38], primarily at the expense of uncharged polar residues. Glutamine in particular seems to be significantly discriminated against in the hyperthermophiles. Although this observation might be rationalized on the basis of

Table 2 Comparison of relative amino acid compositions (in percentages) of mesophiles and thermophiles

	Mesophiles				Thermophiles	
Amino acid	*H. influenzae*	*H. pylori*	*E. coli*	*Synechosystis*	*A. aeolicus*	*M. jannaschii*
A	8.21	6.83	9.55	9.07	5.90	5.54
C	1.03	1.09	1.11	1.01	0.79	1.27
D	4.98	4.77	5.20	5.07	4.32	5.52
E	6.48	6.88	5.91	6.20	9.63	8.67
F	4.46	5.41	3.87	3.75	5.13	4.20
G	6.65	5.76	7.42	7.77	6.75	6.41
H	2.05	2.12	2.26	1.93	1.54	1.43
I	7.10	7.20	5.95	6.31	7.32	10.45
K	6.32	8.94	4.48	4.26	9.40	10.36
L	10.50	11.18	10.56	10.93	10.57	9.38
M	2.44	2.28	2.86	2.12	1.92	2.33
N	4.89	5.83	3.88	3.76	3.60	5.24
P	3.72	3.28	4.41	5.09	4.07	3.38
Q	4.64	3.70	4.42	5.26	2.04	1.44
R	4.47	3.46	5.58	5.18	4.91	3.85
S	5.84	6.81	5.67	5.46	4.79	4.46
T	5.20	4.37	5.35	5.53	4.21	4.06
V	6.68	5.59	7.11	7.10	7.93	6.85
W	1.12	0.70	1.48	1.30	0.93	0.71
Y	3.12	3.68	2.83	2.78	4.13	4.33

	Mesophiles	Thermophiles
Charged residues (DEKRH)	24.11	29.84
Polar/uncharged residues (GSTNQYC)	31.15	26.79
Hydrophobic residues (LMIVWPAF)	44.74	43.36

an increased rate of deamidation of this residue at higher temperatures, aspargine does not appear subject to similar discrimination.

Phylogeny

The placement of the *Aquifex* lineage as one of the earliest divergences in the eubacterial tree[13,14] is interesting because of the insights it could provide into the ancestral eubacterial phenotype, including the hypothesized thermophilic nature of the first bacteria. Protein-based phylogenies often do not support the original rRNA-based placement[15,16,18]. Thus, the availability of some 1,500 genes from an *Aquifex* species would seem to offer a definitive resolution of the phylogeny. However, our analyses of ribosomal proteins, amino-acyl-tRNA synthetases, and other proteins do not do so, showing no consistent picture of the organism's phylogeny. We cannot make a more complete analysis and discussion here, but some observations can be made. These proteins do not yield a statistically significant placement of the *Aquifex* lineage or of other major eubacterial lineages. This situation partially reflects the inadequacy of some protein sequences as indicators of distant molecular genealogy because of their particular evolutionary dynamic, including the patterns and rates of amino-acid replacements. In some cases (such as the aminoacyl-tRNA synthetases for arginine, cysteine, histidine, proline and tyrosine), the analyses are further complicated by the presence of paralogous genes and/or apparent lateral gene transfers. It seems that a more extensive survey of genes and a better sampling of major eubacterial taxa will be required to confidently confirm or refute an early divergence of the *Aquifex* lineage.

Conclusions

Advances in sequencing techniques have allowed us to move beyond studies of single genes to studies of complete genomes only recently[2]. This rapid advance has created the opportunity to begin to characterize an organism with the full knowledge of the genome in hand. The complete genome summarized in this report represents our first view of *A. aeolicus*. The challenge now is to ask specific questions in ways which take advantage of the whole-genome data.

Beyond studies of any single organism in isolation, complete genomes allow comprehensive comparisons between organisms. For instance, comparisons of the similarity of genes can be made that reveal that genes in different categories vary in their relative conservation (Fig. 2). In addition, genome-wide trends are apparent. For example, why is there not more of a tendency to group functionally related genes (for example, biosynthetic pathways) into operons in *A. aeolicus*? This was also seen in the genome sequence of the autotroph *M. jannaschii*[1]. Is this because the autotrophic lifestyle decreases the need for selective regulation? There also seem to be a few multifunctional, fused proteins in *A. aeolicus* and *M. jannaschii*. Although this seems unlikely to be related to autotrophy, it might be associated with extreme thermophily. The large number of diverse genome sequences that will become available in the coming years will allow more detailed correlation of global genomic properties with particular physiologies. □

Methods

Sequencing strategy. The sequencing strategy used to assemble the complete genome was based on the whole genome random (or 'shotgun') approach, which has been successfully used for other genomes of similar size[1–4]. Shotgun sequencing projects are characterized by two phases: an initial completely random phase in which the bulk of the data is collected, followed by a closure phase where directed techniques are used to close gaps and complete the assembly. By pursuing a strategy where only 97% coverage was initially achieved, we were able to limit the number of sequences needed for the random phase to only 10,500 (ref. 39).

Sequences were generated from a small insert library constructed in λ ZAP II vectors[40,41] (average insert length 2.9 kilobase pairs). Two different methods were used for sequencing: first, dye-primer M13-21 and M13 reverse primer (ABI Prism CS+ ready reaction kits, analysed on 48-cm 4% polyacrylamide gels; and second, dye-terminator (ABI Prism FS+) reactions using two pBluescript-specific primers. These reactions were analysed on 36-cm 5% Long-Ranger gels.

The sequence fragments were assembled on an Apple Power Macintosh computer using Sequencher (Gene Codes, Ann Arbor, MI), an assembly and editing program. Assembly was typically performed in batches of roughly 200–400 sequences, and was followed by inspection and editing of the assemblies. All sequences in the set were compared with all others through this process. After assembly, the sequences comprised ~750 contigs at the end of the random phase. Sequences were obtained from both ends of ~200 randomly chosen clones from a fosmid library[42,43]. These sequences were then assembled with consensus sequences derived from the contigs of random-phase sequences using Sequencher. Gaps between contigs were closed by direct sequencing on fosmids not wholly contained within a contig. The fosmid library thus served a purpose analogous to that of the λ-scaffold in other projects[1–4]. The final eight gaps were closed by direct sequencing of polymerase chain reaction (PCR) products generated with the TaqPlus Long PCR System (Stratagene Cloning Systems, La Jolla, CA).

Consequences of reducing the number of sequences in the random phase are the large number of gaps that remain to be closed in the directed phase, and the reduction in overall coverage. To ensure that reduced coverage did not compromise accuracy, ~200 oligonucleotide primers were synthesized to resequence regions of ambiguity identified by visual inspection of the entire assembly. 13,785 sequences, with an average edited read length of 557 base pairs, constitute the final assembly. On the basis of a relatively small number of errors identified during the annotation process, we estimate the error frequency to be <0.01%, comparable to other published genomic sequence estimates.

Gene (ORF + RNA) identification and functional assignment approaches. Coding regions of the *A. aeolicus* genome were analysed and assigned using primarily the programs BLASTP[44] and FASTA[45] to search against a non-redundant protein database. Many analyses were carried out within the context of MAGPIE[46,47], an integrated computing environment for genome analysis. The results of these analyses are available for user interpretation, validation, and categorization. Additional ORFs were identified and start sites refined using the program CRITICA (J. H. Badger and G.J.O., unpublished program). Finally, all presumed 'intergenic regions' were examined with BLASTX for similarities to known protein sequences[48]. Transfer RNA genes were identified with the program tRNAscan-SE[49].

Received 26 August 1997; accepted 3 February 1998.

1. Bult, C. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073 (1996).
2. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–511 (1995).
3. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403 (1995).
4. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388, 539–547 (1997).
5. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24, 4420–4449 (1996).
6. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC7803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3, 109–136 (1996).
7. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1462 (1997).
8. Goffeau, A. *et al.* Life with 6000 genes. *Science* 274, 546 (1996).
9. Huber, R. *et al.* *Aquifex pyrophilus* gen. nov. sp. nov. represents a novel group of marine hyperthermophilic hydrogen oxidizing bacteria. *Arch. Micrtobiol.* 15, 340–351 (1992).
10. Reysenbach, L., Wickham, G. S. & Pace, N. R. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* 60, 2113–2119 (1994).
11. Setchell, W. A. The upper temperature limits of life. *Science* 17, 934–937 (1903).
12. Brock, T. D. The road to Yellowstone—and beyond. *Annu. Rev. Microbiol.* 49, 1–28 (1995).
13. Burggraf, S., Olsen, G. J., Stetter, K. O. & Woese, C. R. A phylogenetic analysis of *Aquifex pyrophilus*. *Syst. Appl. Microbiol.* 15, 353–356 (1992).
14. Pitulle, C. *et al.* Phylogenetic position of the genus *Hydrogenobacter*. *Int. J. Syst. Bacteriol.* 44, 620–626 (1994).
15. Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl Acad. Sci. USA* 93, 7749–7754 (1996).
16. Klenk, H.-P., Palm, P. & Zillig, W. in *Molecular Biology of the Archaea* (eds Pfeifer, F., Palm, P. & Scleeifer, K. H.) 139–147 (Vch Pub, 1994).
17. Bocchetta, M. *et al.* Arrangement and nucleotide sequence of the gene (*fus*) encoding elongation factor G (EF-G) from the hyperthermophilic bacterium *Aquifex pyrophilus*: phylogenetic depth of hyperthermophilic bacteria inferred from analysis of the EF-G/*fus* sequences. *J. Mol. Evol.* 41, 803–812 (1995).
18. Wetmur, J. G. *et al.* Cloning, sequencing, and expression of RecA proteins from three distantly related thermophilic eubacteria. *J. Biol. Chem.* 269, 25928–25935 (1994).
19. Kawasumi, T., Igarashi, Y., Kodama, T. & Minoda, Y. *Hydrogenobacter thermophilus* gen. nov., sp. nov.

[illegible]

22. Riley, M. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952 (1993).
23. Weisburg, W. G., Giovannoni, S. J. & Woese, C. R. The *Deinococcus-Thermus* phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* 11, 128–134 (1989).
24. Beh, M., Strauss, G., Huber, R., Stetter, K. O. & Fuchs, G. Enzymes of the reductive citric acid cycle in the autotrophic eubacterium *Aquifex pyrophilus* and in the archaebacterium *Thermoproteus neutrophilus*. *Arch. Microbiol.* 160, 306–311 (1993).
25. Fuchs, G. in *Autotrophic Bacteria* (eds Schegel, H. G & Bowein, B.) 365–382 (Springer, New York, 1987).
26. Mai, X. & Adams, M. W. Characterization of a fourth type of 2-keto acid-oxidizing enzyme from a hyperthermophilic archaeon: 2-ketoglutarate ferredoxin oxidoreductase from *Thermococcus litoralis*. *J. Bacteriol.* 178, 5890–5896 (1996).
27. Lim, J. H. *et al.* Cloning and expression of superoxide dismutase from *Aquifex pyrophilus*, a hyperthermophilic bacterium. *FEBS Lett.* 406, 142–146 (1997).
28. Bourret, R. B., Borkovich, K. A. & Simon, M. I. Signal transduction pathways involving protein phosphorylation in prokaryotes. *Annu. Rev. Biochem.* 60, 401–441 (1991).
29. Rudolph, J., Tolliday, N., Schmitt, C., Schuster, S. C. & Oesterhelt, D. Phosphorylation in halobacterial signal transduction. *EMBO J.* 14, 4249–4257 (1995).
30. Jarrell, K. F., Bayley, D. P. & Kostyukova, A. S. The archaeal flagellum: a unique motility structure. *J. Bacteriol.* 178, 5057–5064 (1996).
31. Welch, M., Oosawa, K., Aizawa, S. I. & Eisenbach, M. Effects of phosphorylation, Mg^{2+}, and conformation of the chemotaxis protein CheY on its binding to the flagellar switch protein FliM. *Biochemistry* 33, 10470–10467 (1994).
32. Sockett, H., Yamaguchi, S., Kihara, M., Irikura, V. M. & Macnab, R. M. Molecular analysis of the flagellar switch protein FliM of *Salmonella typhimurium*. *J. Bacteriol.* 174, 793–806 (1992).
33. Motoshima, H. *et al.* Molecular cloning and nucleotide sequence of the aminopeptidase T gene of *Thermus aquaticus* YT-1 and its high-level expression in *Escherichia coli*. *Agric. Biol. Chem.* 54, 2385–2392 (1990).
34. Curnow, A. W. *et al.* Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl Acad. Sci. USA* 94, 11819–11826 (1997).
35. Curnow, A. W., Ibba, M. & Söll, D. tRNA-dependent asparagine formation. *Nature* 382, 589–590 (1996).

analysis. *Genomics* 2, 231–239 (1988).

40. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Res.* 16, 7583–7600 (1988).
41. Alting-Mees, M. A. & Short, J. M. pBluescript II: gene mapping vectors. *Nucleic Acids Res.* 17, [illegible] (1989).
42. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* 89, 8794–8797 (1992).
43. Kim, U.-J., Shizuya, H., de Jong, P. J., Birren, B. & Simon, M. I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* 20, 1083–1085 (1992).
44. Altschul, S. F., Fish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990).
45. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* 85, 2444–2448 (1988).
46. Gaasterland, T. & Sensen, C. W. MAGPIE: automated genome interpretation. *Trends Genet.* 12, 76–78 (1996).
47. Gaasterland, T. & Sensen, C. W. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* 78, 302–310 (1996).
48. Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nature Genet.* 3, 266–272 (1993).
49. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964 (1997).

Acknowledgements. This work was supported in part by Department of Energy Microbial Genome Program grants (to R.V.S., C. R. Woese and G.J.O.). We thank C. Woese for his cooperation in the analysis of the genome and interest in the project; K. Stetter for continuing interest; G. Frey, J. Holaska, S. Peralta, D. Hafenbrandl, S. Delk, T. Robinson, and J. Arnett for technical assistance; and D. Robertson, J. Stein, I. Sanyal, T. Richardson, G. Hauska, and K. Williams for discussions.

Correspondence should be addressed to R.V.S. (e-mail: rswanson@diversa.com). Requests for *Aquifex aeolicus* should be addressed to R.H. (e-mail: Robert.huber@biologie.uni-regensburg.de). The sequences have been deposited with GenBank and assigned accession numbers AE000657 (chromosome) and AE000667 (extrachromosomal element).

NCBI **BLAST Search Results** **Entrez** **?**

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= deltaprime.ecoli
(334 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

```
                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

gb|AE000657|AE000657 Aquifex aeolicus complete genome              68    8e-13

 gb|AE000657|AE000657 Aquifex aeolicus complete genome
           Length = 1551335

 Score = 67.5 bits (162), Expect = 8e-13
 Identities = 39/136 (28%), Positives = 58/136 (41%)
 Frame = +1

Query: 25       HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
                HA L     G+G   +   L++ L C+ P   + CG C  C+ +  G  PD   +
Sbjct: 1303996  HAYLFAGPRGVGKTTIARILAKALNCKNPSKGEPCGECENCREIDRGVFPDLIEMDAASN 1304175

Query: 85       KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                +   G+D VR + E +N     G  KV  +                  EEPP  T F L
Sbjct: 1304176  R---GIDDVRALKEAVNYKPIKGKYKVYIIDEAHMLTKEAFNALLKTLEEPPPRTVFVLC 1304346

Query: 145      TREPERLLATLRSRCR 160
                T E +++L T+ SRC+
Sbjct: 1304347  TTEYDKILPTILSRCQ 1304394
```

NCBI **BLAST Search Results** Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

You have searched a database generously provided by the Institute for Genomic Research (TIGR). Their Policy on Early Data Release is:

The Institute for Genomic Research (TIGR) releases data very rapidly to ensure that our scientific colleagues have access to information that may assist them in the search for genes and their biological function. Data releases do not constitute scientific publication, but rather provide investigators with information that may "jump-start" biological experimentation. Users of this information are encouraged to share their results with TIGR in order to improve annotation of the sequence data. Data or information may contain errors or be incomplete and should be regarded as preliminary.

TIGR asks that you acknowledge the source of information obtained from this site in any publication by including the following sentence in both the Materials and Methods and Acknowledgement sections: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" Also include the following text in the Acknowledgements, if applicable: "Sequencing of [organism name] was accomplished with support from [funding agency]." The name of the funding agency for each TIGR project can be found at http://www.tigr.org/tdb/mdb/mdb.html

Similarly, if you display this data or any information derived from it on a Web page, we ask that you prominently display the following notice on that webpage: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" We request that you notify us of your electronic presentation by sending email to www@tigr.org.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= deltaprime.ecoli
(334 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

```
                                                          Score     E
Sequences producing significant alignments:              (bits)  Value
```

```
gb|AE000657|AE000657 Aquifex aeolicus complete genome                    68  1e-12
gb|AE000783|AE000783 Borrelia burgdorferi complete genome                47  2e-06

 gb|AE000657|AE000657 Aquifex aeolicus complete genome
               Length = 1551335

 Score = 67.5 bits (162), Expect = 1e-12
 Identities = 39/136 (28%), Positives = 58/136 (41%)
 Frame = +1

Query: 25      HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
                HA L     G+G    +    L++ L C+ P    + CG C  C+ +  G  PD    +
Sbjct: 1303996 HAYLFAGPRGVGKTTIARILAKALNCKNPSKGEPCGECENCREIDRGVFPDLIEMDAASN 1304175

Query: 85      KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                +   G+D VR + E +N      G  KV  +                    EEPP  T F L
Sbjct: 1304176 R---GIDDVRALKEAVNYKPIKGKYKVYIIDEAHMLTKEAFNALLKTLEEPPPRTVFVLC 1304346

Query: 145     TREPERLLATLRSRCR 160
                T E +++L T+ SRC+
Sbjct: 1304347 TTEYDKILPTILSRCQ 1304394


 Score = 43.0 bits (99), Expect = 4e-05
 Identities = 35/132 (26%), Positives = 56/132 (41%), Gaps = 28/132 (21%)
 Frame = +3

Query: 27      LLIQALPGMGDDALIYALSRYLLCQQ--PQGHKSCGHCRGCQLMQA-------------- 70
                LL      G G      +   ++ +LC++   P G   SC  C+      ++
Sbjct: 1082652 LLFYGKEGSGKTKTAFEFAKGILCKENVPWGCGSCPSCKHVNELEEAFFKGEIEDFKVYK 1082831

Query: 71      ------------GTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXX 118
                            G HPD+   + P      + + ++ +REV          L    KV+ +
Sbjct: 1082832 DKDGKKHFVYLMGEHPDFVVIIPSG--HYIKIEQIREVKNFAYVKPALSRRKVIIIDDAH 1083005

Query: 119     XXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSR 158
                              EEPPA+T  F L T      +L T+ SR
Sbjct: 1083006 AMTSQAANALLKVLEEPPADTTFILTTNRRSAILPTILSR 1083125


 Score = 26.2 bits (56), Expect = 3.9
 Identities = 11/28 (39%), Positives = 15/28 (53%)
 Frame = -3

Query: 32      LPGMGDDALIYALSRYLLCQQPQGHKSC 59
                LPG G+D   +Y L+ Y     +    HK C
Sbjct: 1283214 LPGSGEDFKVYFLTVYRNLTEEHFHKEC 1283131


 Score = 25.1 bits (53), Expect = 8.7
 Identities = 15/45 (33%), Positives = 21/45 (46%)
 Frame = +3

Query: 285    RLQAILGDVCHIREQLMSVTGINRELLITDLLLRIEHYLQPGVVL 329
               R+   +L D   HIR   LM +TGI   +L     + +    H     G  L
Sbjct: 120624 RVAVLLLDRKHIRYFLMDITGIEEKLDFLEPMTTRAHRFHSGGAL 120758


 gb|AE000783|AE000783 Borrelia burgdorferi complete genome
               Length = 910724
```

Following those BLAST hits is the sequence of the <u>contig</u> containing the top hit.

```
TBLASTN 2.0a19MP-WashU [14-Jul-1998] [Build linux-x86 18:51:45 30-Jul-1998]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  statistical significance is estimated under the assumption that the
equivalent of one complete reading frame of the database codes for protein and
that significant alignments will involve only coding reading frames.

Query=  delta prime
        (334 letters)

Database:  /usr/local/db/t_maritima
           948 sequences; 2,352,161 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                     Smallest
                                                                       Sum
                                                     Reading  High  Probability
Sequences producing High-scoring Segment Pairs:        Frame Score  P(N)      N

tm_26                                                    -2    204  3.7e-15   1
tm_804                                                   +3    158  2.2e-10   1
tm_19                                                    -1    133  3.4e-07   1
tm_199                                                   +1     64  0.9999    1


>tm_26
      Length = 18,920

  Minus Strand HSPs:

 Score = 204 (71.8 bits), Expect = 3.7e-15, P = 3.7e-15
 Identities = 56/202 (27%), Positives = 95/202 (47%), Frame = -2

Query:    14 LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
             ++ + Q     H  +     G G   L    L++ L C+  +G + C   CR C+ +   GT
Sbjct:  5536 IIGAIQKNSVAHGYIFAGPRGTGKTTLARILAKSLNCENRKGVEPCNSCRACREIDEGTF 5357

Query:    74 PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLE 133
              D   L   +    N  G+D +R + + +       G  KV  + +  +LT  A NALLKTLE
Sbjct:  5356 MDVIEL--DAASNR-GIDEIRRIRDAVGYRPMEGKYKVYIIDEVHMLTKEAFNALLKTLE 5186

Query:   134 EPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWL-----SREVTMSQDALL 188
             EPP+    F LAT   E++  T+ SRC++        P++      L       +  + ++AL
Sbjct:  5185 EPPSHVVFVLATTNLEKVPPTIISRCQVFEFRNIPDELIEKRLQEVAEAEGIEIDREALS 5006

Query:   189 AALRLSAGSPGAALALFQGDNWQARE 214
               + ++G      AL + +    W+  E
Sbjct:  5005 FIAKRASGGLRDALTMLE-QVWKFSE 4931
```

```
>tm_804
       Length = 1007

  Plus Strand HSPs:

 Score = 158 (55.6 bits), Expect = 2.2e-10, P = 2.2e-10
 Identities = 41/143 (28%), Positives = 65/143 (45%), Frame = +3

Query:    14 LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
             ++ + Q     H  +      G G+  L    L++ L C+    G    C   CR C  +   GT
Sbjct:   249 IIGAIQKNNVAHGYIFAGPRGTGNTTLAIILAKSLNCENRSGVDPCNSCRACIEIDEGTF 428

Query:    74 PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLE 133
              D   L  +   N  G+D +R + + +         G  KV  +     +LT  A NALLK +E
Sbjct:   429 MDVIQL--DAASNR-GIDEIRRIRDAVGYKPMEGKYKVYIID*VHMLTMEAFNALLKAVE 599

Query:   134 EPPAETWFFLATRE----PERLLATL 155
             EPP+    F L T E    P ++++ +
Sbjct:   600 EPPSHVMFVLVTSEL*NGPRKIISNM 677


>tm_19
     Length = 24,312

  Minus Strand HSPs:

 Score = 133 (46.8 bits), Expect = 3.4e-07, P = 3.4e-07
 Identities = 36/97 (37%), Positives = 50/97 (51%), Frame = -1

Query:    75 DYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALLTDAAANALLKTLEE 134
             D   + PE G+N +G+D +R + + LN     L    K V V D    +T  AANA LK LEE
Sbjct: 14943 DVLEIDPE-GEN-IGIDDIRTIKDFLNYSPELYTRKYVIVHDCERMTQQAANAFLKALEE 14770

Query:   135 PPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQY 171
             PP       L TR     LL T++SR       +    P+++
Sbjct: 14769 PPEYAVIVLNTRRWHYLLPTIKSRV-FRVVVNVPKEF 14662


>tm_199
       Length = 1128

  Plus Strand HSPs:

 Score = 64 (22.5 bits), Expect = 8.9, P = 1.00
 Identities = 21/85 (24%), Positives = 40/85 (47%), Frame = +1

Query:   134 EPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWL-----SREVTMSQDALL 188
             EPP+    F LAT   E++   T+ SRC++        P++      L       +   +  ++AL
Sbjct:     1 EPPSHVVFVLATTNLEKVPPTIISRCQVFEFRNIPDELIEKRLQEVAEAEGIEIDREALS 180

Query:   189 AALRLSAGSPGAALALFQGDNWQARE 214
                 +  ++G      AL  +  +     W+   E
Sbjct:   181 FIAKRASGGLRDALTMLE-QVWKFSE 255


Parameters:
  B=5

  ctxfactor=6.00
  E=10
```

The complete genome sequence of the gastric pathogen *Helicobacter pylori*

Jean-F. Tomb*, Owen White*, Anthony R. Kerlavage*, Rebecca A. Clayton*, Granger G. Sutton*, Robert D. Fleischmann*, Karen A. Ketchum*, Hans Peter Klenk*, Steven Gill*, Brian A. Dougherty*, Karen Nelson*, John Quackenbush*, Lixin Zhou*, Ewen F. Kirkness*, Scott Peterson*, Brendan Loftus*, Delwood Richardson*, Robert Dodson*, Hanif G. Khalak*, Anna Glodek*, Keith McKenney*, Lisa M. Fitzegerald*, Norman Lee*, Mark D. Adams*, Erin K. Hickey*, Douglas E. Berg†, Jeanine D. Gocayne*, Teresa R. Utterback*, Jeremy D. Peterson*, Jenny M. Kelley*, Matthew D. Cotton*, Janice M. Weidman*, Claire Fujii*, Cheryl Bowman*, Larry Watthey*, Erik Wallin‡, William S. Hayes§, Mark Borodovsky§, Peter D. Karp‖, Hamilton O. Smith¶, Claire M. Fraser* & J. Craig Venter*

* *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
† *Department of Molecular Biology, School of Medicine, Washington University St Louis, 660 S. Euclid Avenue, St Louis, Missouri 63110, USA*
‡ *Department of Biochemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden*
§ *School of Biology, Georgia Tech, Atlanta, Georgia 30332, USA*
‖ *SRI International, Artificial Intelligence Center, 333 Ravenswood Avenue, Menlo Park, California 94025, USA*
¶ *Department of Molecular Biology and Genetics, School of Medicine, Johns Hopkins University, 725 N. Wolfe Street, Baltimore, Maryland 21205, USA*

Helicobacter pylori, strain 26695, has a circular genome of 1,667,867 base pairs and 1,590 predicted coding sequences. Sequence analysis indicates that _H. pylori_ has well-developed systems for motility, for scavenging iron, and for DNA restriction and modification. Many putative adhesins, lipoproteins and other outer membrane proteins were identified, underscoring the potential complexity of host–pathogen interaction. Based on the large number of sequence-related genes encoding outer membrane proteins and the presence of homopolymeric tracts and dinucleotide repeats in coding sequences, _H. pylori_, like several other mucosal pathogens, probably uses recombination and slipped-strand mispairing within repeats as mechanisms for antigenic variation and adaptive evolution. Consistent with its restricted niche, _H. pylori_ has a few regulatory networks, and a limited metabolic repertoire and biosynthetic capacity. Its survival in acid conditions depends, in part, on its ability to establish a positive inside-membrane potential in low pH.

For most of this century the cause of peptic ulcer disease was thought to be stress-related and the disease to be prevalent in hyperacid producers. The discovery[1] that *Helicobacter pylori* was associated with gastric inflammation and peptic ulcer disease was initially met with scepticism. However, this discovery and subsequent studies on *H. pylori* have revolutionized our view of the gastric environment, the diseases associated with it, and the appropriate treatment regimens[2].

Helicobacter pylori is a micro-aerophilic, Gram-negative, slow-growing, spiral-shaped and flagellated organism. Its most characteristic enzyme is a potent multisubunit urease[3] that is crucial for its survival at acidic pH and for its successful colonization of the gastric environment, a site that few other microbes can colonize[2]. *H. pylori* is probably the most common chronic bacterial infection of humans, present in almost half of the world population[2]. The presence of the bacterium in the gastric mucosa is associated with chronic active gastritis and is implicated in more severe gastric diseases, including chronic atrophic gastritis (a precursor of gastric carcinomas), peptic ulceration and mucosa-associated lymphoid tissue lymphomas[2]. Disease outcome depends on many factors, including bacterial genotype, and host physiology, genotype and dietary habits[4,5]. *H. pylori* infection has also been associated with persistent diarrhoea and increased susceptibility to other infectious diseases[6].

Because of its importance as a human pathogen, our interest in its biology and evolution, and the value of complete genome sequence information for drug discovery and vaccine development, we have

Table 1 Genome features

General	
Coding regions (91.0%)	
Stable RNA (0.7%)	
Non-coding repeats (2.3%)	
Intergenic sequence (6.0%)	
RNA	
Ribosomal RNA	Coordinates
23S-5S	445,306–448,642 bp
23S-5S	1,473,557–1,473,919 bp
16S	1,209,082–1,207,584 bp
16S	1,511,138–1,512,635 bp
5S	448,041–448,618 bp
Transfer RNA	
36 species (7 clusters, 12 single genes)	
Structural RNA	
1 species (ssrD)	629,845–630,124 bp
DNA	
Insertion sequences	
IS605 13 copies (5 full-length, 8 partial)	
IS606 4 copies (2 full-length, 2 partial)	
Distinct G + C regions	Associated genes
region 1 (33% G + C) 452–479 kb	IS605, 5SRNA and repeat 7; *virB4*
region 2 (35% G + C) 539–579 kb	cag PAI (Fig. 4)
region 3 (33% G + C) 1,049–1,071 kb	IS605, 5SRNA and repeat 7
region 4 (43% G + C) 1,264–1,276 kb	β and β′ RNA polymerase, EF-G (fusA)
region 5 (33% G + C) 1,590–1,602 kb	two restriction/modification systems
Coding sequences	
1,590 coding sequences (average 945 bp)	
1,091 identified database match	
499 no database match	

sequenced the genome of a representative *H. pylori* strain by the whole-genome random sequencing method as described for *Haemophilus influenzae*[7], *Mycoplasma genitalium*[8] and *Methanococcus jannaschii*[9].

General features of the genome

Genome analysis. The genome of *H. pylori* strain 26695 consists of a circular chromosome with a size of 1,667,867 base pairs (bp) and average G + C content of 39% (Figs 1 and 2). Five regions within the genome have a significantly different G + C composition (Table 1 and Fig. 1). Two of them contain one or more copies of the insertion sequence IS605 (see below) and are flanked by a 5S ribosomal RNA sequence at one end and a 521 bp repeat (repeat 7) near the other. These two regions are also notable because they contain genes involved in DNA processing and one contains 2 orthologues of the *virB4/ptl* gene, the product of which is required for the transfer of oncogenic T-DNA of *Agrobacterium* and the secretion of the *pertussis* toxin by *Bordetella pertussis*[10]. Another region is the *cag* pathogenicity island (PAI), which is flanked by 31-bp direct repeats, and appears to be the product of lateral transfer[11].

RNA and repeat elements. Thirty-six tRNA species were identified using tRNAscan-SE[12]. These are organized into 7 clusters plus 12 single genes. Two separate sets of 23S–5S and 16S ribosomal RNA (rRNA) genes were identified, along with one orphan 5S gene and one structural RNA gene (Table 1). Associated with each of the two 23S–5S gene clusters is a 6-kilobase (kb) repeat containing a possible operon of 5 ORFs that have no database matches.

Eight repeat families (>97% identity) varying in length from 0.47 to 3.8 kb were found in the chromosome (Figs 1 and 2). Members of repeat 7 are found in intergenic regions, while the others are associated with coding sequences and may represent gene duplications. Repeats 1, 2, 3 and 6 are associated with genes that encode outer-membrane proteins (OMP) (Fig. 3).

Two distinct insertion sequence (IS) elements are present. There are five full-length copies of the previously described IS605[11,13] and two of a newly discovered element designated IS606. In addition, there are eight partial copies of IS605 and two partial copies of IS606. Both elements encode two divergently transcribed transposases (TnpA and TnpB). IS606 has less than 50% nucleotide identity with IS605 and the IS606 transposases have 29% amino-acid identity with their IS605 counterpart. Both copies of the IS606 TnpB may be non-functional owing to frameshifts.

Origin of replication. As a typical eubacterial origin of replication was not identified[14], we arbitrarily designated basepair one at the start of a 7-mer repeat, $(AGTGATT)_{26}$, that produces translational stops in all reading frames, as this repeated DNA is unlikely to contain any coding sequence.

Open reading frames. One thousand five hundred and ninety predicted coding sequences were identified. They were searched against a non-redundant protein database resulting in 1,091 putative identifications that were assigned biological roles using a classification system adapted from Riley[15] (Table 2). The 1,590 predicted genes had an average size of 945 bp, similar to that observed in other prokaryotes[7–9], and no genome-wide strand bias was observed (Fig. 2). More than 70% of the predicted proteins in *H. pylori* have a calculated isoelectric point (p*I*) greater than 7.0, compared to ~40% in *H. influenzae* and *E. coli*. The basic amino acids, arginine and lysine, occur twice as frequently in *H. pylori* proteins as in those of *H. influenzae* and *E. coli*, perhaps reflecting an adaptation of *H. pylori* to gastric acidity.

Paralagous families. Ninety-five paralogous gene families comprising 266 gene products (16% of the total) were identified (www.tigr.org/tdb/mdb/hpdb/hpdb.html). Of these, 67 (173 proteins) have an assigned role. Sixty-four have only 2 members, while the porin/adhesin-like outer membrane protein family (Fig. 2) is the largest with 32 members. The largest number of paralogues with assigned roles fall into the functional categories of cell envelope, transport and binding proteins, and proteins invol in replication. The large number of cell envelope proteins mi reflect either a reduced biosynthetic capacity or a need to adap the challenging gastric environment.

Cell division and protein secretion

The gene content of *H. pylori* suggests that the basic mechanism replication, cell division and secretion are similar to those of *E.* and *H. influenzae*. However, important differences are noted. example, apparently missing from the *H. pylori* genome are orth logues of DnaC, MinC, and the secretory chaperonin, SecB. In or type primosome formation, the DnaB and DnaC proteins form a C complex that delivers the DnaB helicase to the develop primosome complex[16]. The apparent absence of DnaC in *H. py* suggests that either a novel mechanism for recruiting DnaB exists a DnaC orthologue with no detectable sequence similarity present. Similar arguments can be made for other seemingly miss important functions.

H. pylori has a classical set of bacterial chaperones (DnaK, Dn CbpA, GrpE, GroEL, GroES, and HtpG). The transcriptio regulation of *H. pylori* chaperone genes is likely to be differ from that in *E. coli*, as it seems not to have the sigma factors t upregulate chaperone synthesis in *E. coli* (heat-shock sigma 32 a stationary-phase sigma S).

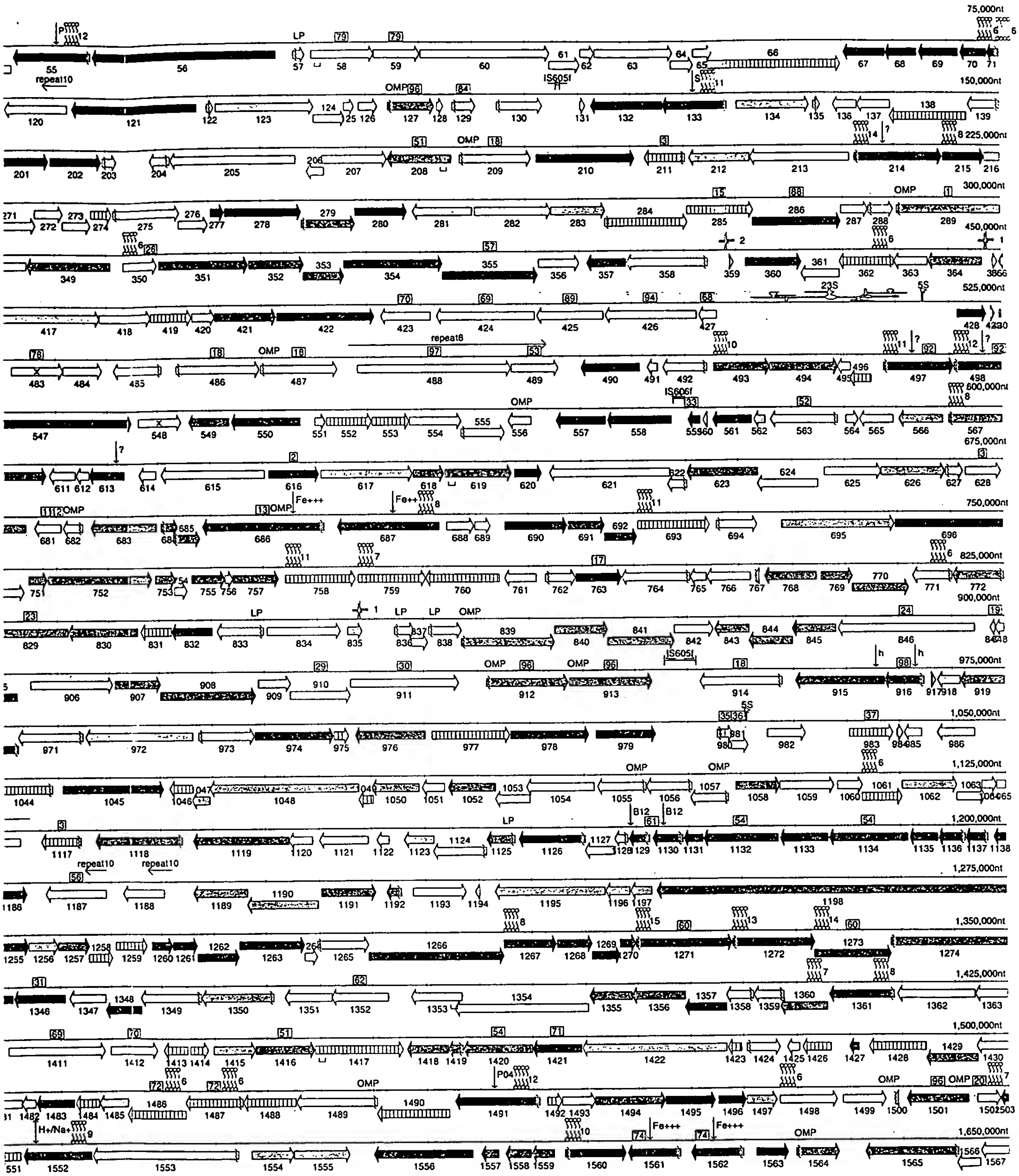
In addition to the SecA-dependent secretory pathway, *H. py* has two specialized export systems. One is associated with the pathogenicity island[11] and the other is the flagellar export path which is assembled from orthologues of FliH, FliI, FliP, FlhA, Fl FliQ, FliR and FliP[17]. Apparently absent from *H. pylori* is a type signal peptidase and orthologues of the dsbABC system, which other species are required for the maturation of pili and pilin-li structures[18] and assembly of surface structures involved in virule and DNA transformation[19].

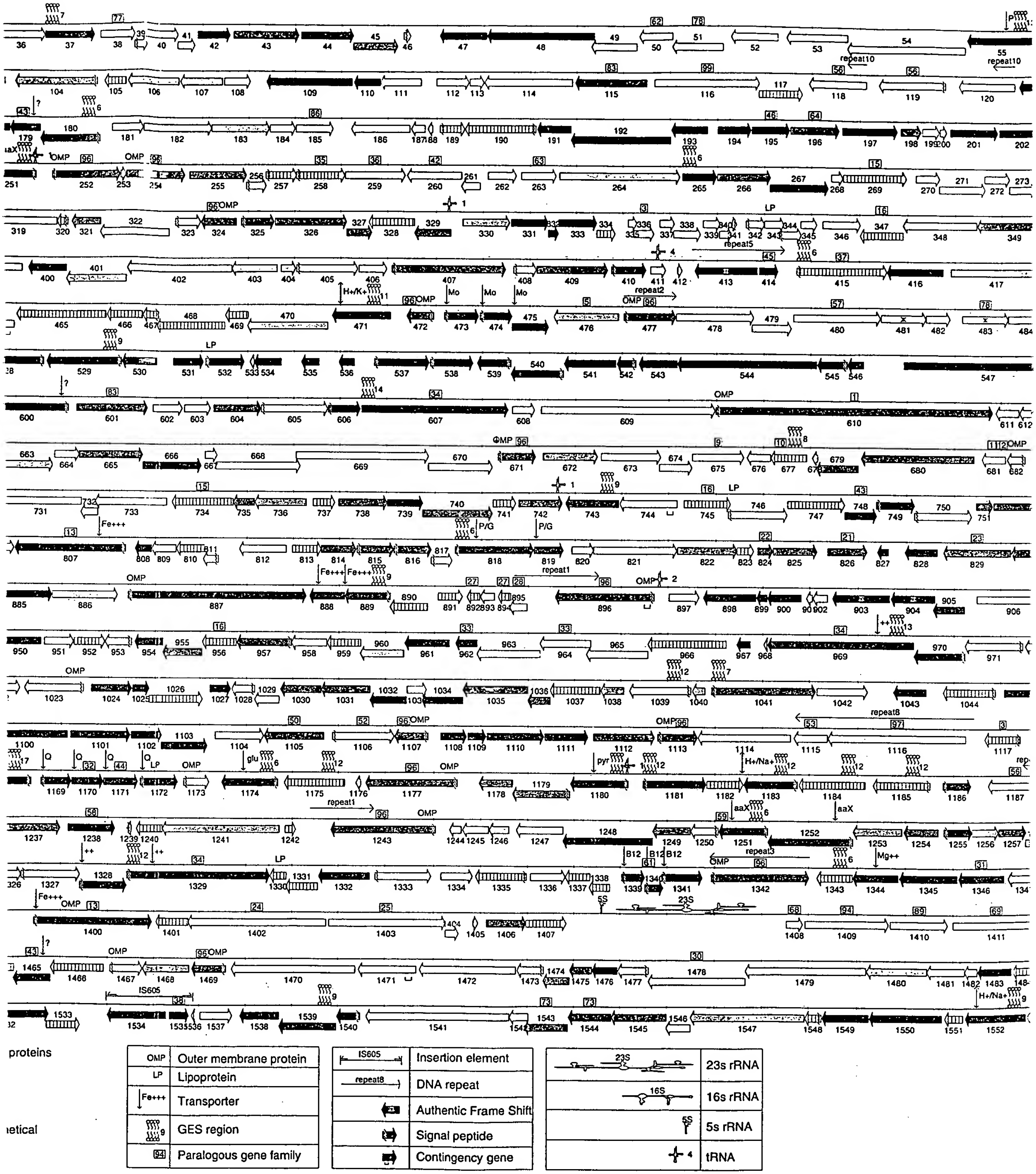
Recombination, repair and restriction systems

Systems for homologous recombination and post-replication, mi match, excision and transcription-coupled repair appear to present in *H. pylori*. Also present are genes with similarity DNA glycosylases which have associated AP endonuclease activi The RecBCD pathway, which mediates homologous recombinati and double-strand break repair, and RecT and RecE orthologu proteins involved in strand exchange during recombination[20], se to be absent. The ability of *H. pylori* to perform mismatch repai suggested by the presence of methyl transferases, mutS and uv However, orthologues of MutH and MutL were not identifi Components of an SOS system also appear to be absent.

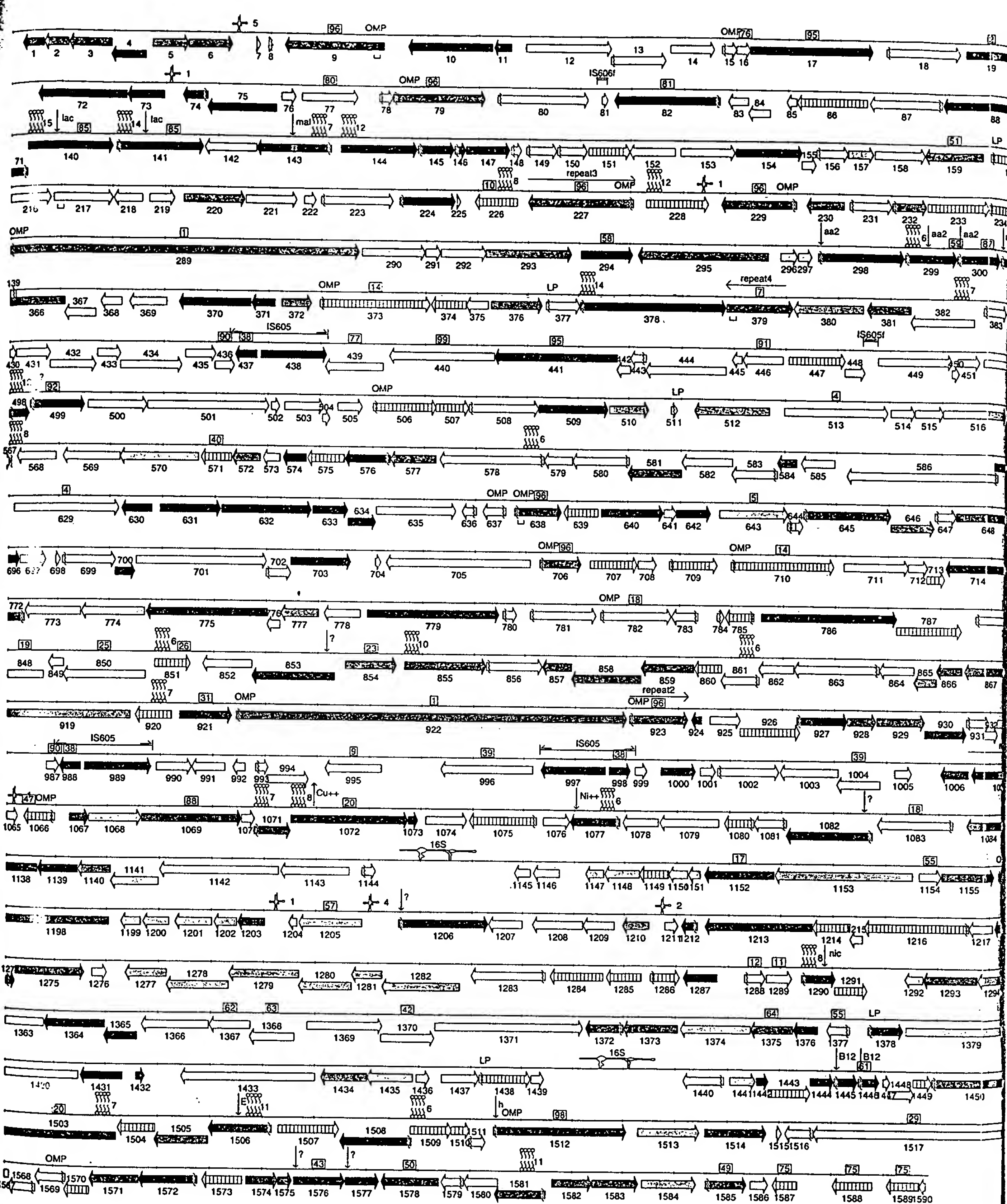
Bacteria commonly use restriction and modification systems degrade foreign DNA. In *H. pylori*, this defence system is developed with eleven restriction-modification systems identifi on the basis of gene order and similarity to endonucleases, meth transferases, and specificity subunits. Three type I, one type II, three type IIS systems were identified, as well as four type systems, including the recently identified epithelial respons

Figure 1 Linear representation of the *H. pylori* 26695 chromosome illustrating location of each predicted protein-coding region, RNA gene, and repeat eleme in the genome. Symbols are as follows: ++, Co^{2+}, Zn^{2+}, Cd^{2+}; ?, unknown; A/G D-alanine/glycine/D-serine; B12, B12/ferric siderophores; E, glutamate; molybdenum; P, proline; P/G, proline/glycine betaine; Q, glutamine serine; a-k, α-ketoglutarate; a/o, arginine/ornithine; aa, amino acids (speci unknown); aa2, dipeptides; aaX, oligopeptides; fum, fumarate, succinate glucose/galactose; h, hemin; lac, L-lactate; mal, malate 2-oxoglutarate nicotinamide mononucleotides; pyr, pyrimidine nucleosides. Numbers ciated with tRNA symbols represent the number of tRNAs at a locus. Num associated with GES represent the number of membrane-spanning dom according to the Goldman, Engelman and Steitz scale as calculated by TopPr


75,000nt
150,000nt
225,000nt
300,000nt
450,000nt
525,000nt
600,000nt
675,000nt
750,000nt
825,000nt
900,000nt
975,000nt
1,050,000nt
1,125,000nt
1,200,000nt
1,275,000nt
1,350,000nt
1,425,000nt
1,500,000nt
1,650,000nt


proteins
ietical
OMP Outer membrane protein
LP Lipoprotein
Fe+++ Transporter
GES region
94 Paralogous gene family
IS605 Insertion element
repeat8 DNA repeat
Authentic Frame Shift
Signal peptide
Contingency gene
23S 23s rRNA
16S 16s rRNA
5S 5s rRNA
4 tRNA


Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid/Phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
Transport/bi
Translation
Transcriptior
Other categ
Conserved h
Unknown
1 kb



volved
might
lapt to

sms of
E. coli
d. For
ortho-
oriC-
n a B-
loping
pylori
ists or
rity is
issing

DnaJ,
tional
fferent
rs that
32 and

pylori
he cag
thway
, FlhB,
ype IV
hich in
in-like
ulence

, mis-
to be
rity to
ctivity.
nation
logues,
, seem
epair is
uvrD,
ntified.

ems to
is well
ntified
nethyl-
II, and
ype III
ponsive

ating the
elements
; A/G/S
ate; Mo
nine;
pecific
ate;
ate;
rs ass
Number
doma
opPred

UST 19

Table 2. List of *H. pylori* genes with putative identifications. Gene numbers correspond to those in Fig. 1. Each identified gene has been assigned a putative role category adapted from ref. 15. Percentages represent per cent identities.

AMINO-ACID BIOSYNTHESIS

General

HP0695	hydantoin utilization protein A (hyuA)	28.6%

Aromatic amino-acid family

HP1038	3-dehydroquinase type II (aroQ)	99.4%
HP0283	3-dehydroquinate synthase (aroB)	38.1%
HP0134	3-deoxy-D-arabino-heptulosonate 7-phosphate synthase (dhs1)	54.6%
HP0401	3-phosphoshikimate 1-carboxyvinyltransferase (aroA)	53.6%
HP1279	anthranilate isomerase (trpC)	47.0%
HP1282	anthranilate synthase component I (trpE)	47.9%
HP1280	anthranilate synthase component II (trpD)	42.5%
HP1281	anthranilate synthase component II (trpD)	40.2%
HP0663	chorismate synthase (aroC)	47.2%
HP1380	prephenate dehydrogenase (tyrA)	30.2%
HP1249	shikimate 5-dehydrogenase (aroE)	36.6%
HP0157	shikimic acid kinase I (aroK)	36.1%
HP1277	tryptophan synthase, alpha subunit (trpA)	46.5%
HP1278	tryptophan synthase, beta subunit (trpB)	66.1%

Aspartate family

HP0649	aspartate ammonia-lyase (aspA)	55.5%
HP1189	aspartate-semialdehyde dehydrogenase (asd)	45.7%
HP1229	aspartokinase (lysC)	48.0%
HP0106	cystathionine gamma-synthase (metB)	47.7%
HP0290	diaminopimelate decarboxylase (dap decarboxylase) (lysA)	42.7%
HP0566	diaminopimelate epimerase (dapF)	30.0%
HP0510	dihydrodipicolinate reductase (dapB)	95.3%
HP1013	dihydrodipicolinate synthetase (dapA)	39.5%
HP0822	homoserine dehydrogenase (metL)	37.7%
HP1050	homoserine kinase (thrB)	27.7%
HP0672	solute-binding signature and mitochondrial signature protein (aspB)	47.3%
HP0212	succinyl-diaminopimelate desuccinylase (dapE)	42.3%
HP0626	tetrahydrodipicolinate N-succinyltransferase (dapD)	36.1%
HP0098	threonine synthase (thrC)	32.9%

Glutamate family

HP0380	glutamate dehydrogenase (gdhA)	59.0%
HP0512	glutamine synthetase (glnA)	48.6%
HP1158	pyrroline-5-carboxylate reductase (proC)	28.9%

Pyruvate family

HP0941	alanine racemase, biosynthetic (alr)	32.4%
HP1468	branched-chain-amino-acid aminotransferase (ilvE)	63.5%
HP0330	ketol-acid reductoisomerase (ilvC)	48.1%

Serine family

HP0107	cysteine synthetase (cysK)	45.7%
HP0096	phosphoglycerate dehydrogenase	31.0%
HP0397	phosphoglycerate dehydrogenase (serA)	32.5%
HP0736	phosphoserine aminotransferase (serC)	30.7%
HP0652	phosphoserine phosphatase (serB)	36.5%
HP1210	serine acetyltransferase (cysE)	98.2%
HP0183	serine hydroxymethyltransferase (glyA)	54.0%

BIOSYNTHESIS OF COFACTORS, PROSTHETIC GROUPS, AND CARRIERS

General

HP0220	synthesis of [Fe-S] cluster (nifS)	48.0%

Biotin

HP0598	8-amino-7-oxononanoate synthase (bioF)	34.9%
HP0976	adenosylmethionine-8-amino-7-oxononanoate aminotransferase (bioA)	49.2%
HP1140	biotin operon repressor/biotin acetyl coenzyme A carboxylase synthetase (birA)	36.9%
HP0407	biotin sulfoxide reductase (bisC)	42.7%
HP1254	biotin synthesis protein (bioC)	32.1%
HP1406	biotin synthetase (bioB)	36.2%
HP0029	dethiobiotin synthetase (bioD)	36.0%

Folic acid

HP1036	7, 8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (folK)	34.6%
HP0587	aminodeoxychorismate lyase (pabC)	32.4%
HP1232	dihydropteroate synthase (folP)	34.5%
HP1545	folylpolyglutamate synthase (folC)	35.2%
HP0928	GTP cyclohydrolase I (folE)	50.9%
HP0577	methylene-tetrahydrofolate dehydrogenase (folD)	48.4%
HP0293	para-aminobenzoate synthetase (pabB)	35.1%

Haem and porphyrin

HP0163	delta-aminolevulinic acid dehydratase (hemB)	50.5%
HP0376	ferrochelatase (hemH)	33.4%
HP0306	glutamate-1-semialdehyde 2,1-aminomutase (hemL)	51.3%
HP0239	glutamyl-tRNA reductase (hemA)	32.7%
HP0665	oxygen-independent coproporphyrinogen III oxidase (hemN)	42.4%
HP1226	oxygen-independent coproporphyrinogen III oxidase (hemN)	37.9%
HP0237	porphobilinogen deaminase (hemC)	45.7%
HP0381	protoporphyrinogen oxidase (hemK)	35.9%
HP0604	uroporphyrinogen decarboxylase (hemE)	46.3%
HP1224	uroporphyrinogen III cosynthase (hemD)	27.6%

Menaquinone and ubiquinone

HP1360	4-hydroxybenzoate octaprenyltransferase (ubiA)	26.6%
HP0929	geranyltranstransferase (ispA)	39.8%
HP0240	octaprenyl-diphosphate synthase (ispB)	31.6%

Molybdopterin

HP0768	molybdenum cofactor biosynthesis protein A (moaA)	31.4%
HP0798	molybdenum cofactor biosynthesis protein C (moaC)	97.9%
HP0172	molybdopterin biosynthesis protein (moeA)	36.3%
HP0755	molybdopterin biosynthesis protein (moeB)	32.2%
HP0799	molybdopterin biosynthesis protein (mog)	50.8%
HP0801	molybdopterin converting factor, subunit 1 (moaD)	31.1%
HP0800	molybdopterin converting factor, subunit 2 (moaE)	31.1%
HP0769	molybdopterin-guanine dinucleotide biosynthesis protein A (mobA)	28.3%

Pantothenate

HP1058	3-methyl-2-oxobutanoate hydroxymethyltransferase (panB)	43.7%
HP0034	aspartate 1-decarboxylase (panD)	50.0%
HP0006	pantoate-beta-alanine ligase (panC)	44.2%
HP0841	pantothenate metabolism flavoprotein (dfp)	31.3%

Pyridoxine

HP1583	pyridoxal phosphate biosynthetic protein A (pdxA)	34.2%
HP1582	pyridoxal phosphate biosynthetic protein J (pdxJ)	42.6%

Riboflavin

HP0802	GTP cyclohydrolase II (ribA)	47.2%
HP0804	GTP cyclohydrolase II/3,4-dihydroxy-2-butanone 4-phosphate synthase (ribA, ribB)	44.0%
HP1505	riboflavin biosynthesis protein (ribG)	33.1%
HP1087	riboflavin biosynthesis regulatory protein (ribC)	28.9%
HP1574	riboflavin synthase alpha subunit (ribC)	32.8%
HP0002	riboflavin synthase beta chain (ribE)	52.4%

Thioredoxin, glutaredoxin and glutathione

HP1118	gamma-glutamyltranspeptidase (ggt)	53.2%
HP1458	thioredoxin	38.3%
HP0824	thioredoxin (trxA)	51.5%
HP1164	thioredoxin reductase (trxB)	28.5%

Thiamine

HP0614	thiamin biosynthesis protein (thiF)	34.6%
HP0843	thiamin phosphate pyrophosphorylase/hyroxyethylthiazole kinase (thiB)	35.7%
HP0845	thiamin phosphate pyrophosphorylase/hyroxyethylthiazole kinase (thiM)	37.9%
HP0844	thiamine biosynthesis protein (thi)	41.0%

Pyridine nucleotides

HP0329	NH(3)-dependent NAD+ synthetase (nadE)	37.5%
HP1355	nicotinate-nucleotide pyrophosphorylase (nadC)	36.3%
HP1356	quinolinate synthetase A (nadA)	34.2%

CELL ENVELOPE

Membranes, lipoproteins and porins

HP1450	60 kDa inner-membrane protein	40.0%
HP0180	apolipoprotein N-acyltransferase (cute)	28.0%
HP0175	cell binding factor 2	34.9%
HP0078	Hypothetical protein	28.4%
HP0567	membrane protein	26.4%
HP1456	membrane-associated lipoprotein (lpp20)	98.9%
HP1564	outer membrane protein	39.9%
HP0009	outer membrane protein (omp1)	0.0%
HP0324	outer membrane protein (omp10)	0.0%
HP0472	outer membrane protein (omp11)	99.5%
HP0477	outer membrane protein (omp12)	0.0%
HP0638	outer membrane protein (omp13)	0.0%
HP0671	outer membrane protein (omp14)	36.0%
HP0706	outer membrane protein (omp15)	33.5%
HP0722	outer membrane protein (omp16)	43.3%
HP0725	outer membrane protein (omp17)	43.3%
HP0796	outer membrane protein (omp18)	0.0%
HP0896	outer membrane protein (omp19)	36.6%
HP0025	outer membrane protein (omp2)	0.0%
HP0912	outer membrane protein (omp20)	0.0%
HP0913	outer membrane protein (omp21)	38.2%
HP0923	outer membrane protein (omp22)	0.0%
HP1107	outer membrane protein (omp23)	0.0%
HP1113	outer membrane protein (omp24)	36.0%
HP1156	outer membrane protein (omp25)	0.0%
HP1157	outer membrane protein (omp26)	23.0%
HP1177	outer membrane protein (omp27)	37.0%
HP1243	outer membrane protein (omp28)	0.0%
HP1342	outer membrane protein (omp29)	0.0%
HP0079	outer membrane protein (omp3)	0.0%
HP1395	outer membrane protein (omp30)	0.0%
HP1469	outer membrane protein (omp31)	0.0%
HP1501	outer membrane protein (omp32)	0.0%
HP0127	outer membrane protein (omp4)	0.0%
HP0227	outer membrane protein (omp5)	36.8%
HP0229	outer membrane protein (omp6)	38.4%
HP0252	outer membrane protein (omp7)	30.6%
HP0254	outer membrane protein (omp8)	37.6%
HP0317	outer membrane protein (omp9)	36.3%
HP0839	outer membrane protein P1 (ompP1)	23.3%
HP0955	prolipoprotein diacylglyceryl transferase (lgt)	34.4%
HP0655	protective surface antigen D15	27.5%
HP1571	rare lipoprotein A (rlpA)	37.6%
HP0610	toxin-like outer membrane protein	26.3%
HP0922	toxin-like outer membrane protein	29.5%
HP0289	toxin-like outer membrane protein	30.6%

Murein sacculus and peptidoglycan

HP0830	amidase	40.6%
HP0738	D-alanine:D-alanine ligase A (ddlA)	28.5%
HP0549	glutamate racemase (glr)	36.6%
HP0772	N-acetylmuramoyl-L-alanine amidase (amiA)	26.8%
HP0597	penicillin-binding protein 1A (PBP-1A)	33.7%
HP1565	penicillin-binding protein 2 (pbp2)	35.0%
HP1125	peptidoglycan associated lipoprotein precursor (omp18)	42.6%
HP0493	phospho-N-acetylmuramoyl-pentapeptide-transferase (mraY)	45.2%
HP0743	rod shape-determining protein (mreB)	37.7%
HP1373	rod shape-determining protein (mreB)	51.9%
HP1372	rod shape-determining protein (mreC)	33.6%
HP0645	soluble lytic murein transglycosylase (slt)	32.2%
HP1543	toxR-activated gene (tagE)	37.2%
HP1544	toxR-activated gene (tagE)	31.2%
HP1155	transferase, peptidoglycan synthesis (murG)	28.2%
HP0740	UDP-MurNac-pentapeptide presynthetase (murF)	25.7%
HP1494	UDP-MurNac-tripeptide synthetase (murE)	36.0%
HP1418	UDP-N-acetylenolpyruvoylglucosamine reductase (murB)	32.7%
HP0648	UDP-N-acetylglucosamine enolpyruvyl transferase (murZ)	46.7%
HP0623	UDP-N-acetylmuramate-alanine ligase (murC)	37.3%
HP0494	UDP-N-acetylmuramoylalanine-D-glutamate ligase (murD)	31.1%

Surface polysaccharides, lipopolysaccharides and antigens

HP0003	3-deoxy-d-manno-octulosonic acid 8-phosphate synthetase (kdsA)	53.4%
HP0957	3-deoxy-d-manno-octulosonic-acid transferase (kdtA)	35.9%
HP0858	ADP-heptose synthase (rfaE)	40.6%
HP1191	ADP-heptose-lps heptosyltransferase II (rfaF)	33.2%
HP0859	ADP-L-glycero-D-mannoheptose-6-epimerase (rfaD)	32.7%
HP0855	alginate O-acetylation protein (algI)	41.8%
HP0326	CMP-N-acetylneuraminic acid synthetase (neuA)	31.9%
HP0230	CTP:CMP-3-deoxy-D-manno-octulosonate-cytidylyl-transferase (kdsB)	36.2%
HP1392	fibronectin/fibrinogen-binding protein	25.7%
HP0379	fucosyltransferase	39.2%
HP0651	fucosyltransferase	39.2%
HP0044	GDP-D-mannose dehydratase (rfbD)	62.1%
HP0867	lipid A disaccharide synthetase (lpxB)	32.0%
HP0159	lipopolysaccharide 1,2-glucosyltransferase (rfaJ)	28.9%
HP0208	lipopolysaccharide 1,2-glucosyltransferase (rfaJ)	26.7%
HP0805	lipooligosaccharide 5G8 epitope biosynthesis-associated protein (lex2B)	36.9%
HP0826	lipooligosaccharide 5G8 epitope biosynthesis-associated protein (lex2B)	39.2%
HP1416	lipopolysaccharide 1,2-glucosyltransferase (rfaJ)	29.2%
HP0679	lipopolysaccharide biosynthesis protein (wbpB)	42.8%
HP1475	lipopolysaccharide core biosynthesis protein (kdtB)	49.0%
HP0279	lipopolysaccharide heptosyltransferase-1 (rfaC)	31.7%
HP0619	lipopolysacharide biosynthesis glycosyl transferase (lic2B)	37.2%
HP1105	LPS biosynthesis protein	28.7%
HP1578	LPS biosynthesis protein	28.1%
HP1581	methicillin resistance protein (llm)	29.2%
HP0857	phosphoheptose isomerase (gmhA)	44.5%
HP1275	phosphomannomutase (algC) [Pseudomonas aeruginosa]	39.6%
HP1429	polysialic acid capsule expression protein (kpsF)	46.0%
HP0366	spore coat polysaccharide biosynthesis protein C	35.3%
HP0178	spore coat polysaccharide biosynthesis protein E	36.2%
HP0421	type 1 capsular polysaccharide biosynthesis protein J (capJ)	29.0%
HP0196	UDP-3-0-(3-hydroxymyristoyl) glucosamine N-acyltransferase (lpxD)	39.5%
HP1052	UDP-3-0-acyl N-acetylglcosamine deacetylase (envA)	44.6%
HP1375	UDP-N-acetylglucosamine acyltransferase (lpxA)	41.8%

Surface structures

HP0840	flaA1 protein	60.2%
HP0325	flagellar basal-body L-ring protein (flgH)	32.7%
HP0351	flagellar basal-body M-ring protein (fliF)	34.4%
HP0246	flagellar basal-body P-ring protein (flgI)	37.9%
HP1557	flagellar basal-body protein (fliE)	37.0%
HP1559	flagellar basal-body rod protein (flgB) (proximal rod protein)	31.0%
HP1558	flagellar basal-body rod protein (flgC) (proximal rod protein)	46.0%
HP1092	flagellar basal-body rod protein (flgG)	35.5%
HP1585	flagellar basal-body rod protein (flgG)	47.7%
HP1041	flagellar biosynthesis protein (flhA)	43.1%
HP1035	flagellar biosynthesis protein (flhF)	35.5%
HP0684	flagellar biosynthesis protein (fliP)	43.4%
HP0770	flagellar biosynthetic protein (flhB)	38.7%
HP0685	flagellar biosynthetic protein (fliP)	55.6%
HP1419	flagellar biosynthetic protein (fliQ)	52.3%
HP0173	flagellar biosynthetic protein (fliR)	26.4%
HP0353	flagellar export protein (fliH)	29.1%
HP1420	flagellar export protein ATP synthase (fliI)	47.6%
HP0870	flagellar hook (flgE)	98.9%
HP0908	flagellar hook (flgE)	30.5%
HP1119	flagellar hook-associated protein 1 (HAP1) (flgK)	27.6%
HP0752	flagellar hook-associated protein 2 (fliD)	28.9%
HP0815	flagellar motor rotation protein (motA)	32.9%
HP0816	flagellar motor rotation protein (motB)	29.7%
HP0352	flagellar motor switch protein (fliG)	37.0%
HP1031	flagellar motor switch protein (fliM)	34.4%
HP0753	flagellar protein (fliS)	32.3%
HP0327	flagellar protein G (flaG)	23.3%
HP0797	flagellar sheath adhesin hpaA	98.5%
HP0584	flagellar switch protein (fliN)	39.7%
HP0601	flagellin A (flaA)	99.8%
HP0115	flagellin B (flaB)	99.0%
HP0295	flagellin B homologue (fla)	32.9%
HP1575	flhB protein (flhB)	40.5%
HP1030	fliY protein (fliY)	29.3%
HP0907	Hook assembly protein, flagella (flgD)	25.5%
HP1274	paralysed flagella protein (pflA)	23.9%
HP0751	polar flagellin (flaG)	21.9%
HP0410	putative neuraminyllactose-binding haemagglutinin homologue (hpaA)	24.2%
HP1192	secreted protein involved in flagellar motility	72.5%
HP1462	secreted protein involved in flagellar motility	96.2%
HP0232	secreted protein involved in flagellar motility	99.2%

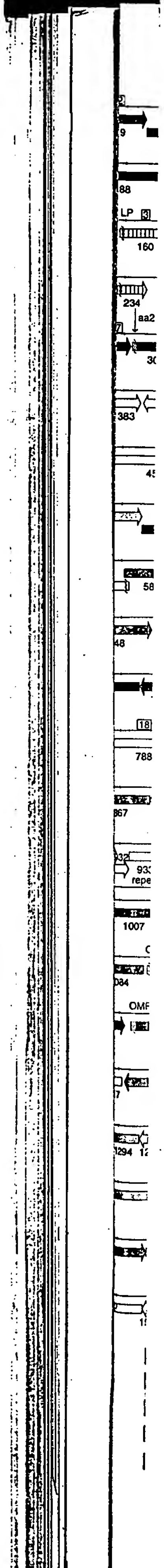
CELLULAR PROCESSES

General

HP0019	chemotaxis protein (cheV)	26.8%
HP0393	chemotaxis protein (cheV)	31.7%
HP0616	chemotaxis protein (cheV)	27.9%
HP1067	chemotaxis protein (cheY)	99.2%
HP0517	GTP-binding protein (era)	95.6%
HP1490	haemolysin	39.2%
HP1086	Haemolysin (tly)	40.2%
HP0599	haemolysin secretion protein precursor (hylB)	45.4%
HP0392	histidine kinase (cheA)	41.4%
HP0099	methyl-accepting chemotaxis protein (tlpA)	32.8%
HP0103	methyl-accepting chemotaxis protein (tlpB)	30.7%
HP0082	methyl-accepting chemotaxis transducer (tlpC)	28.2%
HP0391	purine-binding chemotaxis protein (cheW)	34.3%

Cell division

HP0331	cell division inhibitor (minD)	50.2%
HP0749	cell division membrane protein (ftsX)	25.7%
HP0978	cell division protein (ftsA) protein	31.9%
HP0748	cell division protein (ftsE)	37.6%
HP0286	cell division protein (ftsH)	41.2%
HP1069	cell division protein (ftsH)	98.6%
HP1556	cell division protein (ftsI)	30.6%
HP1090	cell division protein (ftsK)	39.8%
HP1560	cell division protein (ftsW) Escherichia coli	32.7%
HP0763	cell division protein (ftsY)	48.6%



HP0332	cell division topological specificity factor (minE)	33.8%
HP0979	cell divison protein (ftsZ)	43.3%
HP1159	cell filamentation protein (fic)	63.2%
Cell killing		
HP0887	vacuolating cytotoxin	94.7%
Chaperones		
HP0010	chaperone and heat shock protein (groEL)	99.6%
HP0109	chaperone and heat shock protein 70 (dnaK)	63.4%
HP0210	chaperone and heat shock protein C62.5 (htpG)	46.5%
HP0011	co-chaperone (groES)	99.2%
HP1332	co-chaperone and heat-shock protein (dnaJ)	42.7%
HP0110	co-chaperone and heat-shock protein (grpE)	33.0%
HP1024	co-chaperone-curved DNA-binding protein A (CbpA)	37.7%
Chromosome-associated protein		
HP1138	plasmid replication-partition related protein	40.4%
Detoxification		
HP1563	alkyl hydroperoxide reductase (tsaA)	98.5%
HP0875	catalase	99.4%
HP0267	chlorohydrolase	42.6%
HP0243	neutrophil activating protein (napA) (bacterioferritin)	95.8%
HP0389	superoxide dismutase (sodB)	98.6%
HP1452	thiophene and furan oxidizer (tdhF)	37.6%
Protein and peptide secretion		
HP0355	GTP-binding membrane protein (lepA)	57.3%
HP0074	lipoprotein signal peptidase (lspA)	97.0%
HP0786	preprotein translocase subunit (secA)	54.0%
HP1300	preprotein translocase subunit (secY)	41.2%
HP1255	protein translocation protein, low temperature (secG)	30.6%
HP1550	protein-export membrane protein (secD)	38.9%
HP1549	protein-export membrane protein (secF)	35.1%
HP0576	signal peptidase I (lepB)	40.3%
HP1152	signal recognition particle protein (ffh)	41.4%
HP0795	trigger factor (tig)	27.6%
Transformation		
HP0520	cag pathogenicity island protein (cag1)	96.5%
HP0530	cag pathogenicity island protein (cag10)	98.4%
HP0531	cag pathogenicity island protein (cag11)	97.2%
HP0532	cag pathogenicity island protein (cag12)	98.9%
HP0534	cag pathogenicity island protein (cag13)	98.0%
HP0535	cag pathogenicity island protein (cag14)	97.6%
HP0536	cag pathogenicity island protein (cag15)	96.4%
HP0537	cag pathogenicity island protein (cag16)	98.9%
HP0538	cag pathogenicity island protein (cag17)	95.3%
HP0539	cag pathogenicity island protein (cag18)	98.7%
HP0540	cag pathogenicity island protein (cag19)	99.5%
HP0521	cag pathogenicity island protein (cag2)	92.5%
HP0541	cag pathogenicity island protein (cag20)	97.8%
HP0542	cag pathogenicity island protein (cag21)	97.9%
HP0543	cag pathogenicity island protein (cag22)	95.5%
HP0544	cag pathogenicity island protein (cag23)	99.0%
HP0545	cag pathogenicity island protein (cag24)	98.5%
HP0546	cag pathogenicity island protein (cag25)	95.7%
HP0547	cag pathogenicity island protein (cag26)	92.9%
HP0522	cag pathogenicity island protein (cag3)	98.1%
HP0523	cag pathogenicity island protein (cag4)	95.7%
HP0524	cag pathogenicity island protein (cag5)	99.1%
HP0526	cag pathogenicity island protein (cag6)	97.5%
HP0527	cag pathogenicity island protein (cag7)	94.6%
HP0528	cag pathogenicity island protein (cag8)	99.0%
HP0529	cag pathogenicity island protein (cag9)	98.9%
HP1378	competence lipoprotein (comL)	25.5%
HP1361	competence locus E (comE3)	26.7%
HP1006	conjugal transfer protein (traG)	27.3%
HP1421	conjugative transfer regulon protein (trbB)	30.7%
HP0333	DNA processing chain A (dprA)	32.9%
HP0042	trbI protein	31.4%
HP0525	virB11 homologue	100.0%
HP0441	VirB4 homologue	23.5%
HP0017	virB4 homologue (virB4)	25.2%
HP0459	virB4 homologue (virB4)	25.3%
CENTRAL INTERMEDIARY METABOLISM		
General		
HP1014	7-α-hydroxysteroid dehydrogenase (hdhA)	33.2%
HP1186	carbonic anhydrase	37.0%
HP0004	carbonic anhydrase (icfA)	33.3%
HP0869	hydrogenase expression/formation protein (hypA)	28.1%
HP0900	hydrogenase expression/formation protein (hypB)	41.4%
HP0899	hydrogenase expression/formation protein (hypC)	38.5%
HP0898	hydrogenase expression/formation protein (hypD)	47.8%
HP0047	hydrogenase expression/formation protein (hypE)	39.7%
HP0197	S-adenosylmethionine synthetase 2 (metX)	62.1%
Amino sugars		
HP1532	glucosamine fructose-6-phosphate aminotransferase (isomerizing) (glmS)	41.7%
Phosphorus compounds		
HP0620	inorganic pyrophosphatase (ppa)	50.0%
HP0696	N-methylhydantoinase	26.9%
HP1010	polyphosphate kinase (ppk)	38.5%
Polyamine biosynthesis		
HP0422	arginine decarboxylase (speA)	33.3%
HP0020	carboxynorspermidine decarboxylase (nspC)	45.6%
HP0832	spermidine synthase (speE)	26.5%
Other		
HP0070	urease accessory protein (ureE)	97.1%
HP0069	urease accessory protein (ureF)	94.5%
HP0068	urease accessory protein (ureG)	95.0%
HP0067	urease accessory protein (ureH)	96.2%
HP0071	urease accessory protein (ureI)	88.5%
HP0073	urease alpha subunit (ureA) (urea amidohydrolase)	100.0%
HP0072	urease beta subunit (urea amidohydrolase) (ureB)	100.0%
HP0075	urease protein (ureC)	98.0%
ENERGY METABOLISM		
Aerobic		
HP1222	D-lactate dehydrogenase (dld)	27.0%
HP0961	glycerol-3-phosphate dehydrogenase (NAD(P)+)	36.8%
HP0037	NADH-ubiquinone oxidoreductase subunit	19.4%
HP1269	NADH-ubiquinone oxidoreductase, NQO10 subunit (NQO10)	-1.0%
HP1270	NADH-ubiquinone oxidoreductase, NQO11 subunit (NQO11) ((Paracoccus denitrificans)	42.6%
HP1271	NADH-ubiquinone oxidoreductase, NQO12 subunit (NQO12)	43.2%
HP1272	NADH-ubiquinone oxidoreductase, NQO13 subunit (NQO13)	40.2%
HP1273	NADH-ubiquinone oxidoreductase, NQO14 subunit (NQO14)	31.2%
HP1266	NADH-ubiquinone oxidoreductase, NQO3 subunit (NQO3)	31.6%
HP1263	NADH-ubiquinone oxidoreductase, NQO4 subunit (NQO4)(Triticum aestivum)	44.6%
HP1262	NADH-ubiquinone oxidoreductase, NQO5 subunit (NQO5)	-1.0%
HP1261	NADH-ubiquinone oxidoreductase, NQO6 subunit (NQO6)	62.2%
HP1260	NADH-ubiquinone oxidoreductase, NQO7 subunit (NQO7)	40.7%
HP1267	NADH-ubiquinone oxidoreductase, NQO8 subunit (NQO8)	42.4%
HP1268	NADH-ubiquinone oxidoreductase, NQO9 subunit (NQO9)	41.2%
Amino acids and amines		
HP1398	alanine dehydrogenase (ald)	39.6%
HP0294	aliphatic amidase (aimE)	75.4%
HP1238	aliphatic amidase (aimE)	37.2%
HP1399	arginase (rocF)	31.8%
HP0943	D-amino acid dehydrogenase (dadA)	26.2%
HP0056	delta-1-pyrroline-5-carboxylate dehydrogenase (Synechocystis sp.)	32.2%
HP0723	L-asparaginase II (ansB)	64.1%
HP0132	L-serine deaminase (sdaA)	45.8%
Anaerobic		
HP0666	anaerobic glycerol-3-phosphate dehydrogenase, subunit C (glpC)	27.2%
HP0589	ferredoxin oxidoreductase, alpha subunit	42.7%
HP0590	ferredoxin oxidoreductase, beta subunit	43.2%
HP0591	ferredoxin oxidoreductase, gamma subunit	33.3%
HP0193	fumarate reductase, cytochrome b subunit (frdC)	58.8%
HP0192	fumarate reductase, flavoprotein subunit (frdA)	69.4%
HP0191	fumarate reductase, iron-sulfur subunit (frdB)	70.8%
HP1110	pyruvate ferredoxin oxidoreductase, alpha subunit	41.0%
HP1111	pyruvate ferredoxin oxidoreductase, beta subunit	43.7%
HP1109	pyruvate ferredoxin oxidoreductase, delta subunit	47.0%
HP1108	pyruvate ferredoxin oxidoreductase, gamma subunit	37.2%
ATP-protonmotive force interconversion		
HP0828	ATP synthase FO, subunit a (atpB)	37.7%
HP1136	ATP synthase FO, subunit b (atpF)	28.3%
HP1137	ATP synthase FO, subunit bÕ (atpFÕ)	32.5%
HP1212	ATP synthase FO, subunit c (atpE)	41.2%
HP1134	ATP synthase F1, subunit alpha (atpA)	62.7%
HP1132	ATP synthase F1, subunit beta (atpD)	85.6%
HP1135	ATP synthase F1, subunit delta (atpH)	24.6%
HP1131	ATP synthase F1, subunit epsilon (atpC)	32.7%
HP1133	ATP synthase F1, subunit gamma (atpG)	37.8%
Electron transport		
HP0146	cbb3-type cytochrome c oxidase subunit Q (CcoQ)	44.2%
HP0265	cytochrome c biogenesis protein (ccdA)	35.4%
HP0378	cytochrome c biogenesis protein (ycf5)	37.5%
HP0147	cytochrome c oxidase, diheme subunit, membrane-bound (fixP)	33.0%
HP0144	cytochrome c oxidase, heme b and copper-binding subunit, membrane-bound (fixN)	43.9%
HP0145	cytochrome c oxidase, monoheme subunit, membrane-bound (fixO)	45.7%
HP1461	cytochrome c551 peroxidase	48.5%
HP1227	cytochrome c553	38.4%
HP0277	ferredoxin	52.5%
HP0588	ferrodoxin-like protein	42.6%
HP1508	ferrodoxin-like protein	29.4%
HP1161	flavodoxin (fldA)	47.0%
HP0642	NAD(P)H-flavin oxidoreductase	46.1%
HP0954	oxygen-insensitive NAD(P)H nitroreductase	32.7%
HP0634	quinone-reactive Ni/Fe hydrogenase (hydD)	54.7%
HP0633	quinone-reactive Ni/Fe hydrogenase, cytochrome b subunit (hydC)	51.4%
HP0632	quinone-reactive Ni/Fe hydrogenase, large subunit (hydB)	68.5%
HP0631	quinone-reactive Ni/Fe hydrogenase, small subunit (hydA)	68.9%
HP1539	ubiquinol cytochrome c oxidoreductase, cytochrome b subunit (fbcH)	39.3%
HP1538	ubiquinol cytochrome c oxidoreductase, cytochrome c1 subunit (fbcH)	28.8%
HP1540	ubiquinol cytochrome c oxidoreductase, Rieske 2Fe-2S subunit (fbcF)	39.2%
Entner-Doudoroff		
HP1099	2-keto-3-deoxy-6-phosphogluconate aldolase (eda)	50.3%
HP1100	6-phosphogluconate dehydratase	60.7%
Fermentation		
HP0691	3-oxoadipate coA-transferase subunit A (yxjD)	65.5%
HP0692	3-oxoadipate coA-transferase subunit B (yxjE)	73.2%
HP0903	acetate kinase (ackA) (Escherichia coli)	42.3%
HP0904	phosphate acetyltransferase (pta)	51.0%
HP0905	phosphotransacetylase (pta)	26.9%
HP0357	short-chain alcohol dehydrogenase	57.6%
Gluconeogenesis		
HP1385	fructose-1,6-bisphosphatase	36.4%
HP0121	phosphoenolpyruvate synthase (ppsA)	52.4%
HP1345	phosphoglycerate kinase	47.3%
Glycolysis		
HP0154	enolase (eno)	56.9%
HP0176	fructose-bisphosphate aldolase (tsr)	46.0%
HP1103	glucokinase (glk)	41.5%
HP1166	glucose-6-phosphate isomerase (pgi)	53.3%
HP0921	glyceraldehyde-3-phosphate dehydrogenase (gap)	46.5%
HP1346	glyceraldehyde-3-phosphate dehydrogenase (gap)	46.7%
HP0974	phosphoglycerate mutase (pgm)	44.6%
HP0194	triosephosphate isomerase (tpi)	34.5%
Pentose phosphate pathway		
HP1386	D-ribulose-5-phosphate 3 epimerase (rpe)	44.2%
HP1102	glucose-6-phosphate 1-dehydrogenase (devB)	29.2%
HP1101	glucose-6-phosphate dehydrogenase (g6pD)	36.7%
HP1495	transaldolase (tal)	33.5%
HP1088	transketolase A (tktA)	46.7%
HP0354	transketolase B (tktB)	39.7%
Sugars		
HP0574	galactosidase acetyltransferase (lacA)	41.0%
HP0360	UDP-glucose 4-epimerase	43.1%
TCA cycle		
HP0779	aconitase B (acnB)	64.0%
HP0026	citrate synthase (gltA)	47.8%
HP1325	fumarase (fumC)	63.7%
HP0509	glycolate oxidase subunit (glcD)	98.0%
HP0027	isocitrate dehydrogenase (icd)	70.7%
FATTY ACID AND PHOSPHOLIPID METABOLISM		
General		
HP1376	(3R)-hydroxymyristoyl-(acyl carrier protein) dehydratase (fabZ)	47.4%
HP1348	1-acyl-glycerol-3-phosphate acyltransferase (plsC) (Escherichia coli)	32.0%
HP0561	3-ketoacyl-acyl carrier protein reductase (fabG)	45.7%
HP0690	acetyl coenzyme A acetyltransferase (thiolase) (fadA)	52.0%
HP0950	acetyl-CoA carboxylase beta subunit (accD)	49.4%
HP1045	acetyl-CoA synthetase (acoE)	52.3%
HP0557	acetyl-coenzyme A carboxylase (accA)	50.3%
HP0559	acyl carrier protein (acpP)	55.3%
HP0962	acyl carrier protein (acpP)	56.3%
HP0558	beta ketoacyl-acyl carrier protein synthase II (fabF)	50.0%
HP0202	beta-ketoacyl-acyl carrier protein synthase III (fabH)	44.4%
HP0371	biotin carboxyl carrier protein (fabE)	30.8%
HP0370	biotin carboxylase (accC)	52.1%
HP0871	CDP-diglyceride hydrolase (cdh)	73.9%
HP0215	CDP-diglyceride synthetase (cdsA)	42.4%
HP0416	cyclopropane fatty acid synthase (cfa)	39.7%
HP0700	diacylglycerol kinase (dgkA)	45.8%
HP0195	enoyl-(acyl-carrier-protein) reductase (NADH) (fabI)	45.8%
HP0201	fatty acid/phospholipid synthesis protein (plsX)	37.6%
HP0808	Holo-acp synthase (acpS)	29.1%
HP0090	malonyl coenzyme A-acyl carrier protein transacylase (fabD)	35.4%
HP1016	phosphatidylglycerophosphate synthase (pgsA)	35.4%
HP1357	phosphatidylserine decarboxylase proenzyme (psd)	33.2%
HP1071	phosphatidylserine synthase (pssA)	99.6%
HP0499	phospholipase A1 precursor (DR-phospholipase A)	33.8%
PURINES, PYRIMIDINES, NUCLEOSIDES AND NUCLEOTIDES		
General		
HP0757	beta-alanine synthetase homologue	40.0%
2Õ-Deoxyribonucleotide metabolism		
HP0372	deoxycytidine triphosphate deaminase (dcd)	28.2%
HP0865	deoxyuridine 5Õ-triphosphate nucleotidohydrolase (dut)	41.4%
HP0364	ribonucleoside diphosphate reductase, beta subunit (nrdB)	39.0%
HP0680	ribonucleoside-diphosphate reductase 1 alpha subunit (nrdA)	28.4%
HP0825	thioredoxin reductase (trxB)	45.9%
Purine ribonucleotide biosynthesis		
HP0321	5Õ-guanylate kinase (gmk)	44.8%
HP0618	adenylate kinase (adk)	33.3%
HP1112	adenylosuccinate lyase (purB)	49.5%
HP0255	adenylosuccinate synthetase (purA)	44.6%
HP1434	formyltetrahydrofolate hydrolase (purU)	49.1%
HP1218	glycinamide ribonucleotide synthetase (purD)	31.8%
HP0854	GMP reductase (guaC)	31.8%
HP0409	GMP synthase (guaA)	56.1%
HP0829	inosine-5Õ-monophosphate dehydrogenase (guaB)	58.5%
HP0198	nucleoside diphosphate kinase (ndk)	67.7%
HP0742	phosphoribosylpyrophosphate synthetase (prsA)	56.5%
HP1530	purine nucleoside phosphorylase (punB)	20.7%
Pyrimidine ribonucleotide biosynthesis		
HP1084	aspartate transcarbamoylase (pyrB)	38.7%
HP0919	carbamoyl-phosphate synthase (glutamine-hydrolysing) (pyrAb)	48.6%
HP1237	carbamoyl-phosphate synthetase (pyrAa)	39.7%
HP0349	CTP synthetase (pyrG)	50.7%
HP0266	dihydroorotase (pyrC)	-1.0%
HP0581	dihydroorotase (pyrC)	31.5%
HP1011	dihydroorotate dehydrogenase (pyrD)	41.5%
HP1257	orotate phosphoribosyltransferase (pyrE)	35.5%
HP0005	orotidine 5Õ-phosphate decarboxylase (pyrF)	39.0%
HP1474	thymidylate kinase (tmk)	33.9%
HP0777	uridine 5Õ-monophosphate (UMP) kinase (pyrH)	50.4%
Salvage of nucleosides and nucleotides		
HP0104	2Õ,3Õ-cyclic-nucleotide 2Õ-phosphodiesterase (cpdB)	31.8%
HP0572	adenine phosphoribosyltransferase (apt)	50.3%
HP1179	phosphopentomutase (deoB)	55.9%
HP1178	purine-nucleoside phosphorylase (deoD)	55.5%
HP0735	xanthine guanine phosphoribosyl transferase (gpt)	27.1%
Sugar-nucleotide biosynthesis and conversions		
HP0043	mannose-6-phosphate isomerase (pmi) or (algA)	42.8%
HP0045	nodulation protein (nolK)	44.3%
HP0646	UDP-glucose pyrophosphorylase (galU)	65.6%
HP0683	UDP-N-acetylglucosamine pyrophosphorylase (glmU)	40.0%
REGULATORY FUNCTIONS		
General		
HP1032	alternative transcription initiation factor, sigma-F (fliA)	34.6%
HP1168	carbon starvation protein (cstA)	69.8%
HP1442	carbon storage regulator (csrA)	43.3%
HP1027	ferric uptake regulation protein (fur)	39.9%
HP0278	guanosine pentaphosphate phosphohydrolase (gppA)	26.4%
HP0400	penicillin tolerance protein (lytB)	30.6%

l proteins invo
ype proteins mi
a need to adap

asic mechanism
r to those of E.
ices are noted.
genome are ort
onin, SecB. In or
proteins form a
to the develop
DnaC in H. py
ting DnaB exist
ence similarity
seemingly miss

ones (DnaK, Dn
ne transcriptio
ly to be differ
sigma factors t
hock sigma 32 a

pathway, H. pyl
iated with the
ir export pathw
., FliP, FlhA, Flh
pylori is a type
system, which
pili and pilin-li
olved in virulen

ems
-replication, m
air appear to
ith similarity
nuclease activit
is recombinatio
ecE orthologue
mbination[20], see
ismatch repair
mutS and uvr
e not identifie
absent.
cation systems t
e system is we
stems identifie
ucleases, methyl
one type II, an
is four type I
elial responsiv

ome illustrating th
nd repeat elemen
, unknown; A/G/S
E, glutamate; M
Q, glutamine; S
o acids (specific
ate, succinate; gl
2-oxoglutarate; ni
s. Numbers ass
a locus. Number
spanning domain
lated by TopPred

HP0775	penta-phosphate guanosine-3Ó-pyrophosphohydrolase (spoT)	36.7%
HP0224	peptide methionine sulphoxide reductase (msrA)	66.8%
HP1025	putative heat shock protein (hspR)	46.2%
HP1572	regulatory protein DniR	31.9%
HP0703	response regulator	44.2%
HP1021	response regulator	28.7%
HP1043	response regulator	26.8%
HP1365	response regulator	32.4%
HP0166	response regulator (ompR)	51.0%
HP0714	RNA polymerase sigma-54 factor (rpoN)	37.1%
HP0088	RNA polymerase sigma-70 factor (rpoD)	43.5%
HP0792	sigma-54 interacting protein	97.7%
HP0164	signal-transducing protein, histidine kinase	27.1%
HP1364	signal-transducing protein, histidine kinase	24.9%
HP0244	signal-transducing protein, histidine kinase (atoS)	30.0%
HP0048	transcriptional regulator (hypF)	34.5%
HP1287	transcriptional regulator (tenA)	34.7%
HP0727	transcriptional regulator, putative	33.3%

REPLICATION

Degradation of DNA

HP0275	ATP-dependent nuclease (addB)	27.2%
HP0259	exonuclease VII, large subunit (xseA)	37.6%

DNA replication, restriction, modification, recombination and repair

HP0142	A/G-specific adenine glycosylase (mutY)	38.2%
HP0050	adenine specific DNA methyltransferase (dpnA)	37.4%
HP0910	adenine specific DNA methyltransferase (HINDIIM)	33.4%
HP1352	adenine specific DNA methyltransferase (HINFIM)	62.5%
HP0263	adenine specific DNA methyltransferase (hpaim)	33.9%
HP0481	adenine specific DNA methyltransferase (MFOKI)	29.3%
HP0260	adenine specific DNA methyltransferase (mod)	33.9%
HP0593	adenine specific DNA methyltransferase (mod)	38.5%
HP1522	adenine specific DNA methyltransferase (mod)	42.2%
HP0478	adenine specific DNA methyltransferase (VSPIM)	42.1%
HP0054	adenine/cytosine DNA methyltransferase	32.1%
HP0790	anti-codon nuclease masking agent (prrB)	42.9%
HP1529	chromosomal replication initiator protein (dnaA)	34.9%
HP1121	cytosine specific DNA methyltransferase (BSP6IM)	37.0%
HP0051	cytosine specific DNA methyltransferase (DDEM)	39.0%
HP0483	cytosine specific DNA methyltransferase (HPHIMC)	38.7%
HP0701	DNA gyrase, sub A (gyrA)	97.4%
HP0501	DNA gyrase, sub B (gyrB)	46.0%
HP1478	DNA helicase II (uvrD)	35.3%
HP0548	DNA helicase, putative	38.8%
HP0615	DNA ligase (lig)	40.1%
HP0621	DNA mismatch repair protein (MutS)	32.6%
HP1470	DNA polymerase I (polA)	40.0%
HP1460	DNA polymerase III alpha-subunit (dnaE)	42.0%
HP0500	DNA polymerase III beta-subunit (dnaN)	26.0%
HP1231	DNA polymerase III delta prime subunit (holB)	48.6%
HP1387	DNA polymerase III epsilon subunit (dnaQ)	35.1%
HP0717	DNA polymerase III gamma and tau subunits (dnaX)	39.0%
HP0012	DNA primase (dnaG)	36.6%
HP1523	DNA recombinase (recG)	32.7%
HP1393	DNA repair protein (recN)	28.3%
HP0116	DNA topoisomerase I (topA)	45.1%
HP0440	DNA topoisomerase I (topA)	31.7%
HP0602	endonuclease III	36.6%
HP0585	endonuclease III (nth)	40.1%
HP0705	excinuclease ABC subunit A (uvrA)	53.4%
HP1114	excinuclease ABC subunit B (uvrB)	53.1%
HP0821	excinuclease ABC subunit C (uvrC)	31.5%
HP1526	exodeoxyribonuclease (lexA)	58.9%
HP0213	glucose inhibited division protein (gidA)	48.5%
HP1063	glucose-inhibited division protein (gidB)	32.9%
HP1553	helicase	33.0%
HP0883	Holliday junction DNA helicase (ruvA)	39.0%
HP1059	Holliday junction DNA helicase (ruvB)	54.6%
HP0877	Holliday junction endodeoxyribonuclease (ruvC)	34.7%
HP0675	integrase/recombinase (xerC)	31.8%
HP0995	integrase/recombinase (xerD)	27.8%
HP0323	membrane bound endonuclease (nuc)	31.1%
HP0676	methylated-DNAÑprotein-cysteine methyltransferase (dat1)	41.0%
HP0387	primosomal protein replication factor (priA)	36.3%
HP0153	recombinase (recA)	99.1%
HP0925	recombinational DNA repair protein (recR)	36.5%
HP0911	rep helicase, single-stranded DNA-dependent ATPase (rep)	33.8%
HP1362	replicative DNA helicase (dnaB)	39.4%
HP1383	restriction modification system S subunit	38.1%
HP0661	ribonuclease H (rnhA)	58.4%
HP1323	ribonuclease HII (rnhB)	36.3%
HP1245	single-strand DNA-binding protein (ssb)	32.6%
HP0348	single-stranded-DNA-specific exonuclease (recJ)	33.6%
HP1009	site-specific recombinase	21.3%
HP1541	transcription-repair coupling factor (trcF)	37.7%
HP0462	type I restriction enzyme S protein (hsdS)	37.0%
HP0463	type I restriction enzyme M protein (hsdM)	29.4%
HP0464	type I restriction enzyme R protein (hsdR)	31.7%
HP0846	type I restriction enzyme R protein (hsdR)	48.0%
HP0848	type I restriction enzyme S protein (hsdS)	37.0%
HP0850	type I restriction enzyme M protein (hsdM)	54.4%
HP1402	type I restriction enzyme R protein (hsdR)	26.6%
HP1403	type I restriction enzyme M protein (hsdM)	37.1%
HP1404	type I restriction enzyme S protein (hsdS)	36.0%
HP0092	type II restriction enzyme M protein (hsdM)	55.3%
HP0091	type II restriction enzyme R protein (hsdR)	50.7%
HP1369	type III restriction enzyme M protein (mod)	45.6%
HP1370	type III restriction enzyme M protein (mod)	37.0%
HP1371	type III restriction enzyme R protein	26.2%
HP0592	type III restriction enzyme R protein (res)	30.6%
HP1521	type III restriction enzyme R protein (res)	33.1%
HP1472	type IIS restriction enzyme M protein (mod)	32.4%
HP1367	type IIS restriction enzyme M1 protein (mod) ((Moraxella bovis)	59.3%
HP1368	type IIS restriction enzyme M2 protein (mod)	33.0%
HP1517	type IIS restriction enzyme R and M protein (ECO57IR)	26.7%
HP1471	type IIS restriction enzyme R protein (BCGIB)	28.2%
HP1366	type IIS restriction enzyme R protein (MBOIIR)	37.1%
HP1208	ulcer associated adenine specific DNA methyltransferase	93.4%
HP1209	ulcer-associated gene restriction endonuclease (iceA)	95.5%
HP1347	uracil-DNA glycosylase (ung)	43.1%

TRANSCRIPTION

Degradation of RNA

HP1213	polynucleotide phosphorylase (pnp)	38.9%
	DNA-dependent RNA polymerase	
HP1293	DNA-directed RNA polymerase, alpha subunit (rpoA)	35.3%
HP1198	DNA-directed RNA polymerase, beta subunit (rpoB)	47.8%

Transcription factors

HP0866	transcription elongation factor GreA (greA)	50.3%
HP1514	transcription termination factor NusA (nusA)	39.1%
HP0001	transcription termination factor NusB (nusB)	30.2%
HP1203	transcription termination factor NusG (nusG)	41.0%
HP0550	transcription termination factor Rho (rho)	56.6%

RNA processing

HP0640	poly(A) polymerase (papS)	37.4%
HP0662	ribonuclease III (rnc)	37.3%

TRANSLATION

General

HP0944	translation initiation inhibitor, putative	45.6%

Aminoacyl tRNA synthetases

HP1241	alanyl-tRNA synthetase (alaS)	44.9%
HP0319	arginyl-tRNA synthetase (argS)	35.8%
HP0617	aspartyl-tRNA synthetase (aspS)	50.1%
HP0886	cysteinyl-tRNA synthetase (cysS)	97.3%
HP0476	glutamyl-tRNA synthetase (gltX)	43.1%
HP0643	glutamyl-tRNA synthetase (gltX)	39.8%
HP0960	glycyl-tRNA synthetase, alpha subunit (glyQ)	60.1%
HP0972	glycyl-tRNA synthetase, beta subunit (glyS)	33.6%
HP1190	histidyl-tRNA synthetase (hisS)	32.4%
HP1422	isoleucyl-tRNA synthetase (ileS)	49.7%
HP1547	leucyl-tRNA synthetase (leuS)	45.9%
HP0182	lysyl-tRNA synthetase (lysS)	58.6%
HP0417	methionyl-tRNA synthetase (metS)	42.4%
HP0403	phenylalanyl-tRNA synthetase, alpha subunit (pheS)	48.7%
HP0402	phenylalanyl-tRNA synthetase, beta subunit (pheT)	30.0%
HP0238	prolyl-tRNA synthetase (proS)	39.8%
HP1480	seryl-tRNA synthetase (serS)	48.3%
HP0123	threonyl-tRNA synthetase (thrS)	42.1%
HP1253	tryptophanyl-tRNA synthetase (trpS)	52.6%
HP0774	tyrosyl-tRNA synthetase (tyrS)	54.7%
HP1153	valyl-tRNA synthetase (valS)	43.7%

Degradation of proteins, peptides and glycopeptides

HP0570	aminopeptidase a/i (pepA)	38.5%
HP0033	ATP-dependent C1p protease (clpA)	40.3%
HP0794	ATP-dependent clp protease proteolytic component (clpP)	64.6%
HP1379	ATP-dependent protease (lon)	43.9%
HP0223	ATP-dependent protease (sms)	41.0%
HP1374	ATP-dependent protease ATPase subunit (clpX)	56.3%
HP0264	ATP-dependent protease binding subunit (clpB)	97.7%
HP0169	collagenase (prtC)	40.1%
HP0516	heat-shock protein (hslU) ORF1	98.4%
HP0515	heat-shock protein (hslV)	57.1%
HP0470	oligoendopeptidase F (pepF)	97.9%
HP0657	processing protease (ymxG)	24.2%
HP1485	proline dipeptidase (pepQ)	35.2%
HP1350	protease	40.6%
HP1012	protease (pqqE)	29.6%
HP1435	protease IV (PspA)	41.7%
HP0404	protein kinase C inhibitor (SP:P16436)	40.2%
HP1019	serine protease (htrA)	52.9%
HP1584	sialoglycoprotease (gcp)	35.7%
HP0382	zinc-metalloprotease (YJR117W)	36.2%

Nucleoproteins

HP0835	histone-like DNA-binding protein HU (hup)	44.6%

Protein modification

HP0363	L-isoaspartyl-protein carboxyl methyltransferase (pcm)	43.0%
HP1299	methionine amino peptidase (map)	43.0%
HP1441	peptidyl-prolyl *cis-trans* isomerase B, cyclosporin-type rotamase (ppi)	58.1%
HP1123	peptidyl-prolyl *cis-trans* isomerase, FKBP-type rotamase (slyD)	40.4%
HP0793	polypeptide deformylase (def)	41.8%

Ribosomal proteins: synthesis and modification

HP1201	ribosomal protein L1 (rpl1)	52.0%
HP1200	ribosomal protein L10 (rpl10)	30.4%
HP1202	ribosomal protein L11 (rpl11)	63.8%
HP1068	ribosomal protein L11 methyltransferase (prmA)	38.4%
HP0084	ribosomal protein L13 (rpl13)	50.0%
HP1309	ribosomal protein L14 (rpl14)	65.9%
HP1301	ribosomal protein L15 (rpl15)	42.5%
HP1312	ribosomal protein L16 (rpl16)	62.4%
HP1292	ribosomal protein L17 (rpl17)	48.3%
HP1303	ribosomal protein L18 (rpl18)	45.5%
HP1147	ribosomal protein L19 (rpl19)	60.9%
HP1316	ribosomal protein L2 (rpl2)	58.9%
HP0126	ribosomal protein L20 (rpl20)	54.8%
HP0296	ribosomal protein L21 (rpl21)	46.1%
HP1314	ribosomal protein L22 (rpl22)	44.9%
HP1317	ribosomal protein L23 (rpl23)	31.7%
HP1308	ribosomal protein L24 (rpl24)	52.2%
HP0297	ribosomal protein L27 (rpl27)	64.7%
HP0491	ribosomal protein L28 (rpL28)	41.7%
HP1311	ribosomal protein L29 (rpL29)	45.6%
HP1319	ribosomal protein L3 (rpl3)	41.8%
HP0551	ribosomal protein L31 (rpl31)	49.3%
HP0200	ribosomal protein L32 (rpl32)	41.7%
HP1204	ribosomal protein L33 (rpL33)	55.1%
HP1447	ribosomal protein L34 (rpl34)	70.5%
HP0125	ribosomal protein L35 (rpl35)	60.8%
HP1297	ribosomal protein L36 (rpl36)	81.6%
HP1318	ribosomal protein L4 (rpl4)	40.6%
HP1307	ribosomal protein L5 (rpl5)	53.1%
HP1304	ribosomal protein L6 (rpl6)	44.4%
HP1199	ribosomal protein L7/L12 (rpl7/l12)	65.0%
HP0514	ribosomal protein L9 (rpl9)	39.6%
HP0399	ribosomal protein S1 (rps1)	30.5%
HP1320	ribosomal protein S10 (rps10)	58.2%
HP1295	ribosomal protein S11 (rps11)	56.2%
HP1197	ribosomal protein S12 (rps12)	79.0%
HP1296	ribosomal protein S13 (rps13)	55.8%
HP1306	ribosomal protein S14 (rpS14)	68.3%
HP1040	ribosomal protein S15 (rps15)	57.8%
HP1151	ribosomal protein S16 (rpS16)	46.8%
HP1310	ribosomal protein S17 (rps17)	55.4%
HP1244	ribosomal protein S18 (rps18)	55.2%
HP1315	ribosomal protein S19 (rps19)	61.1%
HP1554	ribosomal protein S2 (rps2)	49.6%
HP0076	ribosomal protein S20 (rps20)	41.4%
HP0562	ribosomal protein S21 (rps21)	42.4%
HP1313	ribosomal protein S3 (rps3)	56.7%
HP1294	ribosomal protein S4 (rps4)	51.2%
HP1302	ribosomal protein S5 (rps5)	65.5%
HP1246	ribosomal protein S6 (rps6)	32.1%
HP1196	ribosomal protein S7 (rps7)	62.2%
HP1305	ribosomal protein S8 (rps8)	45.0%
HP0083	ribosomal protein S9 (rps9)	50.4%
HP1047	ribosome-binding factor A (rbfA)	26.3%

tRNA modification

HP1141	methionyl-tRNA formyltransferase (fmt)	37.5%
HP1497	peptidyl-tRNA hydrolase (pth)	46.6%
HP0361	pseudouridylate synthase I (hisT)	32.2%
HP1448	ribonuclease P, protein component (rnpA)	29.3%
HP1062	S-adenosylmethionine:tRNA ribosyltransferase-isomerase (queA)	39.3%
HP1513	selenocystein synthase (selA)	36.2%
HP1148	tRNA (guanine-N1)-methyltransferase (trmD)	39.1%
HP1415	tRNA delta(2)-isopentenylpyrophosphate transferase (miaA)	30.7%
HP0281	tRNA-guanine transglycosylase (tgt)	45.6%

Translation factors

HP0247	ATP-dependent RNA helicase, DEAD-box family (deaD)	37.7%
HP0077	peptide chain release factor RF-1 (prfA)	52.6%
HP0171	peptide chain release factor RF-2 (prfB)	49.6%
HP1256	ribosome releasing factor (frr)	43.7%
HP1195	translation elongation factor EF-G (fusA)	67.5%
HP0177	translation elongation factor EF-P (efp)	45.1%
HP1555	translation elongation factor EF-Ts (tsf)	43.1%
HP1205	translation elongation factor EF-Tu (tufB)	89.5%
HP1298	translation initiation factor EF-1 (infA)	65.3%
HP1048	translation initiation factor IF-2 (infB)	45.4%
HP0124	translation initiation factor IF-3 (infC)	43.4%

TRANSPORT AND BINDING PROTEINS

General

HP0179	ABC transporter, ATP-binding protein	66.7%
HP0613	ABC transporter, ATP-binding protein	31.1%
HP0715	ABC transporter, ATP-binding protein	52.3%
HP1576	ABC transporter, ATP-binding protein (abc)	48.2%
HP1465	ABC transporter, ATP-binding protein (HI1087)	37.8%
HP1220	ABC transporter, ATP-binding protein (yhcG)	31.5%
HP0853	ABC transporter, ATP-binding protein (yheS)	36.3%
HP1577	ABC transporter, permease protein (yaeE)	43.1%
HP0607	acriflavine resistance protein (acrB)	29.7%
HP1432	histidine and glutamine-rich protein	50.0%
HP1427	histidine-rich, metal binding polypeptide (hpn)	100.0%
HP1206	multidrug-resistance protein (hetA)	26.2%
HP1082	multidrug-resistance protein (msbA)	32.4%
HP0600	multidrug-resistance protein (spaB)	29.7%
HP1181	multidrug-efflux transporter	29.1%
HP0497	sodium- and chloride-dependent transporter	31.7%
HP0498	sodium- and chloride-dependent transporter	30.8%
HP0214	sodium-dependent transporter (huNaDC-1)	36.6%

Amino acids, peptides and amines

HP0940	amino acid ABC transporter, periplasmic binding protein (yckK)	41.5%
HP0939	amino acid ABC transporter, permease protein (yckJ)	46.9%
HP1017	amino acid permease (rocE)	41.7%
HP0942	D-alanine glycine permease (dagA)	44.5%
HP0301	dipeptide ABC transporter, ATP-binding protein (dppD)	59.5%
HP0302	dipeptide ABC transporter, ATP-binding protein (dppF)	54.8%
HP0298	dipeptide ABC transporter, periplasmic dipeptide-binding protein (dppA)	39.8%
HP0299	dipeptide ABC transporter, permease protein (dppB)	49.3%
HP0300	dipeptide ABC transporter, permease protein (dppC)	52.5%
HP1506	glutamate permease (gltS)	56.9%
HP1171	glutamine ABC transporter, ATP-binding protein (glnQ)	51.9%
HP1172	glutamine ABC transporter, periplasmic glutamine-binding protein (glnH)	32.2%
HP1169	glutamine ABC transporter, permease protein (glnP)	27.6%
HP1170	glutamine ABC transporter, permease protein (glnP)	30.9%
HP0250	oligopeptide ABC transporter, ATP-binding protein (oppD)	39.1%
HP1252	oligopeptide ABC transporter, periplasmic oligopeptide-binding protein (oppA)	28.7%
HP1251	oligopeptide ABC transporter, permease protein (oppB)	59.6%
HP0251	oligopeptide ABC transporter, permease protein (oppC)	31.4%
HP0819	osmoprotection protein (proV)	38.3%
HP0818	osmoprotection protein (proWX)	30.4%
HP0055	proline permease (putP)	51.4%
HP0936	proline/betaine transporter (proP)	29.1%
HP0133	serine transporter (sdaC)	44.6%

Anions

HP0475	molybdenum ABC transporter, ATP-binding protein (modD)	38.4%
HP0473	molybdenum ABC transporter, periplasmic molybdate-binding protein (modA)	95.9%
HP0474	molybdenum ABC transporter, permease protein (modB)	28.7%
HP0313	nitrite extrusion protein (narK)	23.6%
HP1491	phosphate permease	34.8%

Carbohydrates, organic alcohols and acids

HP0143	2-oxoglutarate/malate translocator (SODIT1)	37.0%
HP1091	alpha-ketoglutarate permease (kgtP)	45.9%
HP0724	anaerobic C4-dicarboxylate transport protein (dcuA)	53.8%
HP1174	glucose/galactose transporter (gluP)	53.6%
HP0141	L-lactate permease (lctP)	55.5%
HP0140	L-lactate permease (lctP)	58.7%

Cations

HP0791	cadmium-transporting ATPase, P-type (cadA)	97.5%
HP0969	cation efflux system protein (czcA)	37.3%
HP1328	cation efflux system protein (czcA)	28.9%
HP1329	cation efflux system protein (czcA)	31.3%
HP1503	cation-transporting ATPase, P-type (copA)	30.3%
HP1073	copper ion binding protein (copP)	92.4%
HP1072	copper-transporting ATPase, P-type (copA)	93.9%
HP0471	glutathione-regulated potassium-efflux system protein (kefB)	99.3%
HP0687	iron(II) transport protein (feoB)	33.6%
HP1561	iron(III) ABC transporter, periplasmic iron-binding protein (ceuE)	27.5%
HP1562	iron(III) ABC transporter, periplasmic iron-binding protein (ceuE)	28.2%
HP0888	iron(III) dicitrate ABC transporter, ATP-binding protein (fecE)	34.4%
HP0889	iron(III) dicitrate ABC transporter, permease protein (fecD)	38.3%
HP0686	iron(III) dicitrate transport protein (fecA)	29.7%
HP0807	iron(III) dicitrate transport protein (fecA)	28.5%
HP1400	iron(III) dicitrate transport protein (fecA)	26.3%
HP1344	magnesium and cobalt transport protein (corA)	26.3%
HP1183	NA+/H+ antiporter (napA)	26.6%
HP1552	Na+/H+ antiporter (nhaA)	49.2%
HP1077	nickel transport protein (nixA)	98.7%
HP0490	putative potassium channel protein, putative	25.7%

Nucleosides, purines and pyrimidines

HP1290	nicotinamide mononucleotide transporter (pnuC)	28.0%
HP1180	pyrimidine nucleoside transport protein (nupC)	32.9%

Other

HP0876	iron-regulated outer membrane protein (frpB)	27.6%
HP0915	iron-regulated outer membrane protein (frpB)	28.1%
HP0916	iron-regulated outer membrane protein (frpB)	28.8%
HP1129	biopolymer transport protein (exbD)	29.7%
HP1130	biopolymer transport protein (exbB)	33.5%
HP1339	biopolymer transport protein (exbB)	46.8%
HP1340	biopolymer transport protein (exbD)	35.8%
HP1445	biopolymer transport protein (exbB)	45.5%
HP1446	biopolymer transport protein (exbD)	36.2%
HP1512	iron-regulated outer membrane protein (frpB)	26.6%
HP0653	nonheme iron-containing ferritin (pfr)	99.4%
HP1341	siderophore-mediated iron transport protein (tonB)	37.2%

OTHER CATEGORIES

General

HP0924	4-oxalocrotonate tautomerase (dmpl)	37.7%
HP1034	ATP-binding protein (ylxH)	36.3%
HP1000	PARA protein	29.7%
HP1139	SpoOJ regulator (soj)	47.4%
HP0827	ss-DNA binding protein 12RNP2 precursor	46.8%

Adaptations and atypical conditions

HP1496	general stress protein (ctc)	26.5%
HP1483	gerC2 protein (gerC2)	33.3%
HP0927	heat shock protein (htpX)	32.8%
HP0280	heat shock protein B (ibpB)	27.2%
HP1228	invasion protein (invA)	38.2%
HP0970	nickel-cobalt-cadmium resistance protein (nccB)	21.1%
HP1444	small protein (smpB)	42.1%
HP0930	stationary-phase survival protein (surE)	37.7%
HP0315	virulence associated protein D (vapD)	70.2%
HP0967	virulence associated protein D (vapD)	28.9%
HP1248	virulence associated protein homolog (vacB)	36.0%
HP0885	virulence factor mviN protein (mviN)	33.5%

Colicin-related functions

HP1126	colicin tolerance-like protein (tolB)	25.7%
HP0428	phage/colicin/tellurite resistance cluster terY protein	25.6%

Drug and analog sensitivity

HP1431	16S rRNA (adenosine-N6,N6-)-dimethyltransferase (ksgA)	35.5%
HP0606	membrane fusion protein (mtrC)	24.2%
HP0630	modulator of drug activity (mda66)	62.3%
HP1476	phenylacrylic acid decarboxylase	39.7%
HP1165	tetracycline resistance protein tetA(P), putative	27.0%

Transposon-related functions

HP1008	IS200 insertion sequence from SARA17	33.9%
HP0414	IS200 insertion sequence from SARA17	33.9%
HP0988	IS605 transposase (tnpA)	97.2%
HP0998	IS605 transposase (tnpA)	97.2%
HP1096	IS605 transposase (tnpA)	97.2%
HP1535	IS605 transposase (tnpA)	97.2%
HP0437	IS605 transposase (tnpA)	97.2%
HP0989	IS605 transposase (tnpB)	93.4%
HP0997	IS605 transposase (tnpB)	93.4%
HP1095	IS605 transposase (tnpB)	93.4%
HP1534	IS605 transposase (tnpB)	93.4%
HP0438	IS605 transposase (tnpB)	93.4%
HP0413	transposase-like protein, PS3IS	33.6%
HP1007	transposase-like protein, PS3IS	34.3%

Other

HP0739	2-hydroxy-6-oxohepta-2,4-dienoate hydrolase	30.1%

HYPOTHETICAL

General

HP0831	conserved hypothetical ATP binding protein	32.3%
HP0066	conserved hypothetical ATP-binding protein	34.7%
HP0269	conserved hypothetical ATP-binding protein	37.7%
HP0312	conserved hypothetical ATP-binding protein	34.1%
HP1321	conserved hypothetical ATP-binding protein	30.8%
HP1430	conserved hypothetical ATP-binding protein	38.1%
HP1507	conserved hypothetical ATP-binding protein	51.6%
HP1567	conserved hypothetical ATP-binding protein	40.9%
HP1026	conserved hypothetical helicase-like protein	35.2%
HP0022	conserved hypothetical integral membrane protein	30.8%
HP0189	conserved hypothetical integral membrane protein	43.1%
HP0226	conserved hypothetical integral membrane protein	27.6%
HP0228	conserved hypothetical integral membrane protein	43.2%
HP0234	conserved hypothetical integral membrane protein	32.4%
HP0258	conserved hypothetical integral membrane protein	32.7%
HP0284	conserved hypothetical integral membrane protein	29.2%
HP0362	conserved hypothetical integral membrane protein	28.8%
HP0415	conserved hypothetical integral membrane protein	44.4%
HP0467	conserved hypothetical integral membrane protein	100.0%
HP0571	conserved hypothetical integral membrane protein	29.5%
HP0644	conserved hypothetical integral membrane protein	30.3%
HP0677	conserved hypothetical integral membrane protein	28.5%
HP0693	conserved hypothetical integral membrane protein	46.7%
HP0718	conserved hypothetical integral membrane protein	33.5%
HP0737	conserved hypothetical integral membrane protein	33.3%
HP0758	conserved hypothetical integral membrane protein	47.6%
HP0759	conserved hypothetical integral membrane protein	31.1%
HP0787	conserved hypothetical integral membrane protein	25.2%
HP0851	conserved hypothetical integral membrane protein	37.3%
HP0920	conserved hypothetical integral membrane protein	36.3%
HP0946	conserved hypothetical integral membrane protein	35.9%
HP0952	conserved hypothetical integral membrane protein	38.5%
HP0983	conserved hypothetical integral membrane protein	32.8%
HP1044	conserved hypothetical integral membrane protein	30.6%
HP1061	conserved hypothetical integral membrane protein	35.0%
HP1080	conserved hypothetical integral membrane protein	44.0%
HP1162	conserved hypothetical integral membrane protein	27.6%
HP1175	conserved hypothetical integral membrane protein	40.6%
HP1184	conserved hypothetical integral membrane protein	23.5%
HP1185	conserved hypothetical integral membrane protein	55.5%
HP1225	conserved hypothetical integral membrane protein	31.6%
HP1234	conserved hypothetical integral membrane protein	29.0%
HP1235	conserved hypothetical integral membrane protein	30.9%
HP1330	conserved hypothetical integral membrane protein	41.7%
HP1331	conserved hypothetical integral membrane protein	33.6%
HP1343	conserved hypothetical integral membrane protein	49.1%
HP1363	conserved hypothetical integral membrane protein	33.1%
HP1407	conserved hypothetical integral membrane protein	22.4%
HP1466	conserved hypothetical integral membrane protein	30.9%
HP1484	conserved hypothetical integral membrane protein	41.2%
HP1486	conserved hypothetical integral membrane protein	23.8%
HP1487	conserved hypothetical integral membrane protein	30.7%
HP1509	conserved hypothetical integral membrane protein	34.3%
HP1548	conserved hypothetical integral membrane protein	30.6%
HP0138	conserved hypothetical iron-sulfur protein	41.2%
HP1438	conserved hypothetical lipoprotein	32.0%
HP0151	conserved hypothetical membrane protein	21.8%
HP0575	conserved hypothetical membrane protein	38.8%
HP1258	conserved hypothetical mitochondrial protein 4	23.2%
HP1492	conserved hypothetical nifU-like protein	48.2%
HP0032	conserved hypothetical protein	37.0%
HP0035	conserved hypothetical protein	34.1%
HP0086	conserved hypothetical protein	28.7%
HP0094	conserved hypothetical protein	29.8%
HP0100	conserved hypothetical protein	32.0%
HP0102	conserved hypothetical protein	29.3%
HP0105	conserved hypothetical protein	39.7%
HP0117	conserved hypothetical protein	34.2%
HP0162	conserved hypothetical protein	36.7%
HP0216	conserved hypothetical protein	33.9%
HP0233	conserved hypothetical protein	30.5%
HP0248	conserved hypothetical protein	30.7%
HP0274	conserved hypothetical protein	38.5%
HP0285	conserved hypothetical protein	30.8%
HP0309	conserved hypothetical protein	31.3%
HP0310	conserved hypothetical protein	33.7%
HP0318	conserved hypothetical protein	47.2%
HP0328	conserved hypothetical protein	30.7%
HP0334	conserved hypothetical protein	30.8%
HP0347	conserved hypothetical protein	31.8%
HP0373	conserved hypothetical protein	31.4%
HP0374	conserved hypothetical protein	24.7%
HP0388	conserved hypothetical protein	39.8%
HP0395	conserved hypothetical protein	39.9%
HP0396	conserved hypothetical protein	33.7%
HP0419	conserved hypothetical protein	45.6%
HP0447	conserved hypothetical protein	38.2%
HP0465	conserved hypothetical protein	95.5%
HP0466	conserved hypothetical protein	95.7%
HP0468	conserved hypothetical protein	97.1%
HP0469	conserved hypothetical protein	95.1%
HP0496	conserved hypothetical protein	99.2%
HP0507	conserved hypothetical protein	37.2%
HP0519	conserved hypothetical protein	95.3%
HP0552	conserved hypothetical protein	37.8%
HP0553	conserved hypothetical protein	30.0%
HP0639	conserved hypothetical protein	41.0%
HP0654	conserved hypothetical protein	32.0%
HP0658	conserved hypothetical protein	36.0%
HP0707	conserved hypothetical protein	40.1%
HP0709	conserved hypothetical protein	49.6%
HP0710	conserved hypothetical protein	33.7%
HP0716	conserved hypothetical protein	30.2%
HP0728	conserved hypothetical protein	[illegible]
HP0734	conserved hypothetical protein	[illegible]
HP0741	conserved hypothetical protein	[illegible]
HP0745	conserved hypothetical protein	[illegible]
HP0747	conserved hypothetical protein	32.4%
HP0760	conserved hypothetical protein	36.1%
HP0810	conserved hypothetical protein	31.0%
HP0813	conserved hypothetical protein	32.5%
HP0823	conserved hypothetical protein	27.8%
HP0860	conserved hypothetical protein	52.1%
HP0890	conserved hypothetical protein	32.2%
HP0891	conserved hypothetical protein	33.6%
HP0892	conserved hypothetical protein	39.1%
HP0894	conserved hypothetical protein	39.8%
HP0926	conserved hypothetical protein	30.7%
HP0934	conserved hypothetical protein	33.6%
HP0956	conserved hypothetical protein	36.2%
HP0959	conserved hypothetical protein	[illegible]
HP0966	conserved hypothetical protein	[illegible]
HP0975	conserved hypothetical protein	[illegible]
HP1020	conserved hypothetical protein	31.5%
HP1037	conserved hypothetical protein	95.9%
HP1046	conserved hypothetical protein	32.6%
HP1049	conserved hypothetical protein	39.7%
HP1066	conserved hypothetical protein	41.3%
HP1149	conserved hypothetical protein	24.7%
HP1160	conserved hypothetical protein	34.7%
HP1182	conserved hypothetical protein	34.6%
HP1214	conserved hypothetical protein	21.5%
HP1221	conserved hypothetical protein	42.4%
HP1240	conserved hypothetical protein	22.5%
HP1242	conserved hypothetical protein	42.3%
HP1259	conserved hypothetical protein	44.6%
HP1284	conserved hypothetical protein	36.8%
HP1291	conserved hypothetical protein	26.3%
HP1335	conserved hypothetical protein	33.9%
HP1337	conserved hypothetical protein	[illegible]7.2%
HP1338	conserved hypothetical protein	36.2%
HP1394	conserved hypothetical protein	33.6%
HP1401	conserved hypothetical protein	27.5%
HP1413	conserved hypothetical protein	41.6%
HP1414	conserved hypothetical protein	27.4%
HP1417	conserved hypothetical protein	23.7%
HP1423	conserved hypothetical protein	40.3%
HP1426	conserved hypothetical protein	40.0%
HP1428	conserved hypothetical protein	37.8%
HP1443	conserved hypothetical protein	37.9%
HP1449	conserved hypothetical protein	39.0%
HP1453	conserved hypothetical protein	26.8%
HP1459	conserved hypothetical protein	30.1%
HP1504	conserved hypothetical protein	23.9%
HP1510	conserved hypothetical protein	30.6%
HP1533	conserved hypothetical protein	25.4%
HP1570	conserved hypothetical protein	40.5%
HP1573	conserved hypothetical protein	42.2%
HP1587	conserved hypothetical protein	39.0%
HP1588	conserved hypothetical protein	32.0%
HP1589	conserved hypothetical protein	35.1%
HP0713	conserved hypothetical protein (plasmid pHPM180)	41.8%
HP0028	conserved hypothetical secreted protein	42.1%
HP0139	conserved hypothetical secreted protein	37.1%
HP0160	conserved hypothetical secreted protein	30.6%
HP0190	conserved hypothetical secreted protein	31.4%
HP0211	conserved hypothetical secreted protein	24.3%
HP0235	conserved hypothetical secreted protein	31.5%
HP0257	conserved hypothetical secreted protein	29.2%
HP0320	conserved hypothetical secreted protein	36.4%
HP0506	conserved hypothetical secreted protein	29.8%
HP0518	conserved hypothetical secreted protein	95.9%
HP0785	conserved hypothetical secreted protein	26.6%
HP0949	conserved hypothetical secreted protein	39.7%
HP0977	conserved hypothetical secreted protein	29.4%
HP0980	conserved hypothetical secreted protein	57.4%
HP1075	conserved hypothetical secreted protein	42.9%
HP1098	conserved hypothetical secreted protein	27.0%
HP1117	conserved hypothetical secreted protein	32.3%
HP1216	conserved hypothetical secreted protein	31.9%
HP1285	conserved hypothetical secreted protein	38.0%
HP1286	conserved hypothetical secreted protein	37.5%
HP1464	conserved hypothetical secreted protein	27.4%
HP1488	conserved hypothetical secreted protein	29.8%
HP1551	conserved hypothetical secreted protein	42.7%

UNKNOWN

General

HP0390	adhesin-thiol peroxidase (tagD)	38.3%
HP1193	aldo-keto reductase, putative	46.6%
HP0872	alkylphosphonate uptake protein (phnA)	61.1%
HP0207	ATP-binding protein (mpr)	38.9%
HP0136	bacterioferritin comigratory protein (bcp)	35.5%
HP0485	catalase-like protein	30.8%
HP1104	cinnamyl-alcohol dehydrogenase ELI3-2 (cad)	44.0%
HP0981	exonuclease VII-like protein (xseA)	42.5%
HP0569	GTP-binding protein (gtp1)	48.1%
HP0303	GTP-binding protein (obg)	48.2%
HP0834	GTP-binding protein homologue (yphC)	36.7%
HP0480	GTP-binding protein, fusA-homolog (yihK)	54.1%
HP1489	lipase-like protein	21.7%
HP0405	nifS-like protein	27.3%
HP0221	nifU-like protein	37.3%
HP0658	PET112-like protein	45.4%
HP0089	pfs protein (pfs)	34.5%
HP0322	poly E-rich protein	26.7%
HP0625	protein E (gcpE)	[illegible]
HP0431	protein phosphatase 2C homolog (ptc1)	30.7%
HP0624	solute-binding signature and mitochondrial signature protein (aspB)	26.4%
HP0377	thiol:disulfide interchange protein (dsbC), putative	26.4%

endonuclease, *iceA1*, and its associated DNA adenine methyltransferase (M. HypI) genes[21,22]. In addition to the complete systems, seven adenine-specific, and four cytosine-specific methyltransferases, and one of unknown specificity were found. Each of these has an adjacent gene with no database match, suggesting that they may function as part of restriction-modification systems.

Transcription and translation

Although analysis of gene content suggests that *H. pylori* has a basic transcriptional and translational machinery similar to that of *E. coli*, interesting differences are observed. For example, no genes for a catalytic activity in tRNA maturation (*rnd*, *rph*, or *rnpB*) were identified and of the three known ribonucleases involved in mRNA degradation, only polyribonucleotide phosphorylase was found. Twenty-one genes coding for 18 of the 20 tRNA synthetases normally required for protein biosynthesis were found.

As in most other completely sequenced bacterial genomes, the gene for glutaminyl-tRNA synthetase, *glnS*, is missing, and the existence of a transamidation process is assumed. It is also possible that the product of the second glutamyl-tRNA synthetase gene, *gltX*, present in *H. pylori*, may have acquired the glutaminyl-tRNA synthetase function. *H. pylori* provides the first example of a bacterial genome apparently lacking an asparaginyl-tRNA synthetase gene, *asnS*. A transamidation process to form *Asn-tRNAAsn* from *Asp-tRNAAsn* has been reported for the archaeon *Haloferax volcanii*[22] and may also operate in *H. pylori*. Most intriguing, however, is the finding that in *H. pylori* the genes encoding the β and β′ subunits of RNA polymerase are fused. In all studied prokaryotes the two genes are contiguous, but separate, and are part of the same transcriptional unit. Whether this gene fusion in *H. pylori* results in a fused protein, or whether the transcriptional or translational product of the fusion is subject to splicing, is currently not known. It is worth noting that an artificial fusion of the *E. coli* *rpoB* and *rpoC* genes is viable and results in a transcriptional complex, which has the same stoichiometry as the native complex (K. Severinov, personal communication).

Adhesion and adaptive antigenic variation

Most pathogens show tropism to specific tissues or cell types and often use several adherence mechanisms for successful attachment. *H. pylori* may use at least five different adhesins to attach to gastric epithelial cells[5]. One of them, HpaA (HP0797), was previously identified as a lipoprotein in the flagellar sheath and outer membrane[5,23]. In addition to the HpaA orthologue, we have identified 19 other lipoproteins. Few have an identifiable function, but some are likely to contribute to the adherence capacity of the organism.

Two adhesins[24–26], one of which mediates attachment to the Lewis^b histo-blood group antigens, belong to the large family of outer membrane proteins (OMP) (Fig. 3) (T. Boren and R. Haas, personal communication). It is conceivable that other members of these closely related proteins also act as adhesins. Given the large number of sequence-related genes encoding putative surface-exposed proteins, the potential exists for recombinational events leading to mosaic organization. This could be the basis for antigenic variation in *H. pylori* and an effective mechanism for host defence evasion, as seen in *M. genitalium*[27].

At least one other mechanism for antigenic variation could operate in *H. pylori*. The DNA sequence at the beginning of eight genes, including five members of the OMP family, contain stretches of CT or AG dinucleotide repeats (Table 3a). In addition, poly(C) or poly(G) tracts occur within the coding sequence of nine other genes (Table 3b). Slipped-strand mispairing within such repeats are documented features of one mechanism of genotypic variation[28,29]. These mechanisms may have evolved in bacterial pathogens to increase the frequency of phenotypic variation in genes involved in

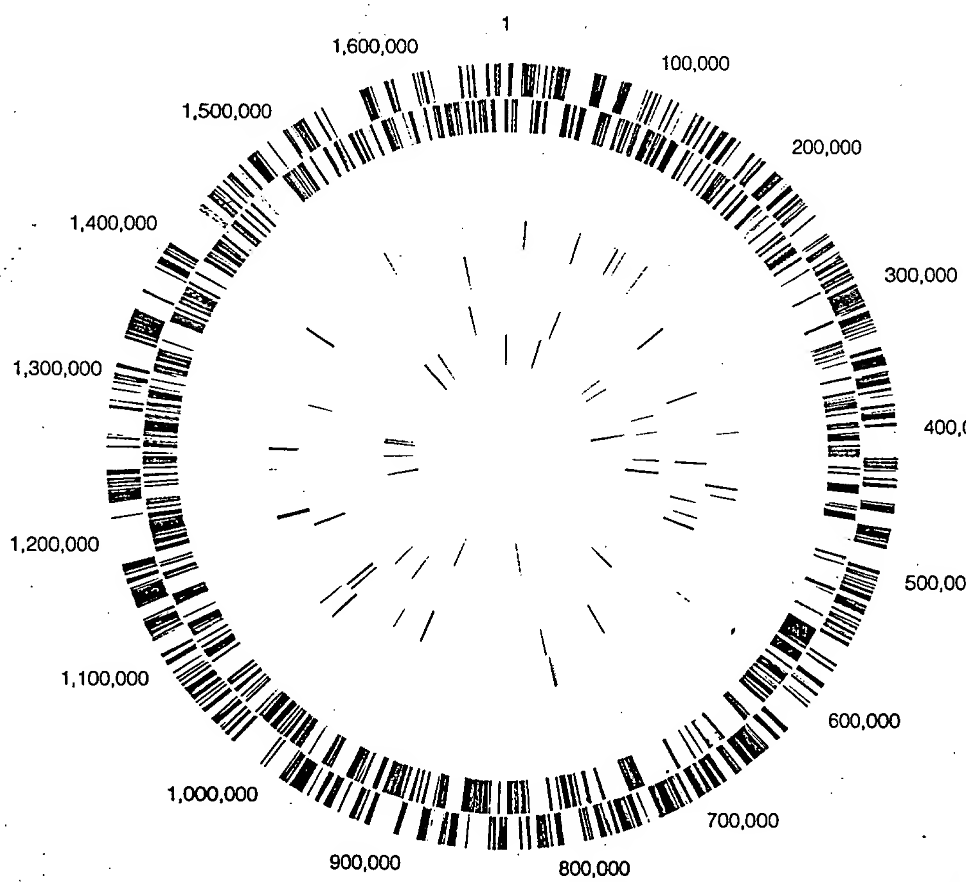


Figure 2 Circular representation of the *H. pylori* 26695 chromosome. Outer concentric circle: predicted coding regions on the plus strand classified as to role according to the colour code in Fig. 1 (except for unknowns and hypotheticals, which are in black). Second concentric circle: predicted coding regions on the minus strand. Third and fourth concentric circles: IS elements (red) and other repeats (green) on the plus and minus strand, respectively. Fifth and sixth concentric circles: tRNAs (blue), rRNAs (red), and sRNAs (green) on the plus and minus strand, respectively.

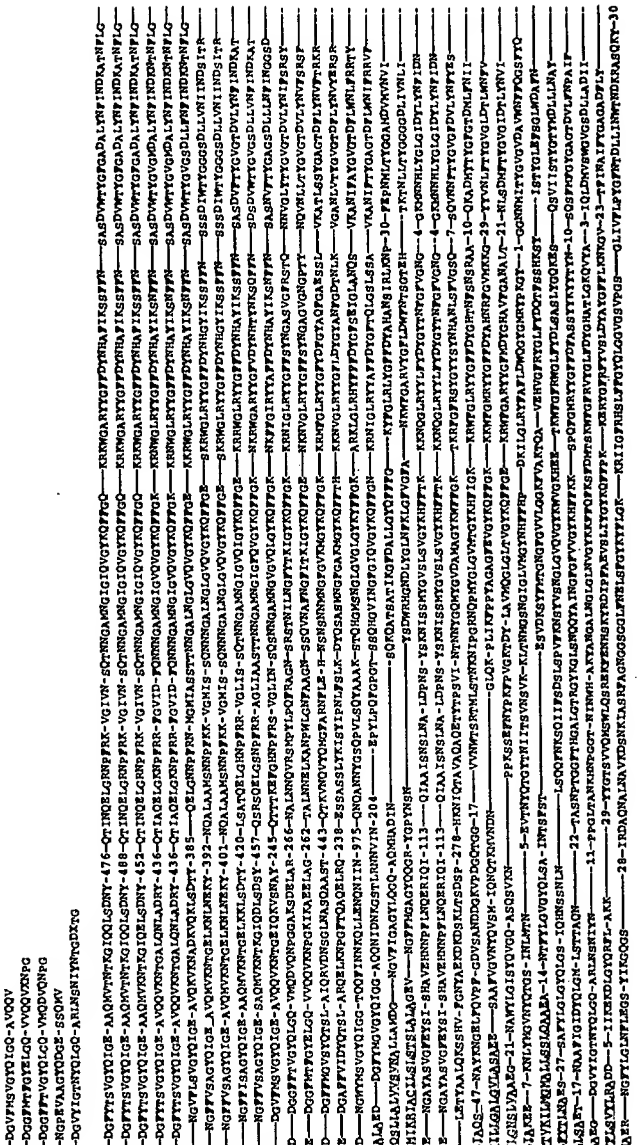


Figure 3 Multiple sequence alignment of members of the outer membrane protein family of *H. pylori*. These proteins were identified as OMPs based on the characteristic alternating hydrophobic residues at their carboxy termini. All members of this family have one domain of similarity at the amino-terminal end and seven domains of similarity at their carboxy-terminal end. Note that the first 11 of these OMPs share extensive similarity over their entire length. Four of the OMPs were identified as porins (Hops) based on identity to published amino-terminal sequences, represented at the top of the alignment[50]. The most likely candidate for HopD is HP0913, which has 15 matches to the first 20-residue N-terminal peptide sequence[50]. These differences may be due to strain variability. The program Signal-P[48] was used to identify cleavage sites and signal peptides (underlined). Four of the OMPs have TTG start codons (HP1156, HP0252, HP1113, HP0796). Numbers embedded in the sequences represent amino acids omitted from the alignment. The star symbols indicate that HP722, HP725 and HP9 proteins contain a frameshift in their signal-peptide-coding region. These frameshifts are associated with the presence of dinucleotide repeats (Table 3).

ical interactions with their hosts[28]. Such 'contingency' genes encode surface structures like pilins, lipoproteins or enzymes that produce lipopolysaccharide molecules[28]. Our analysis suggests that the seventeen genes reported in Table 3a,b belong to this category and thus may provide an example of adaptive evolution in *H. pylori*. Phenotypic variation at the transcriptional level may also operate in *H. pylori*. Examples of repetitive DNA mediating transcriptional control have been documented by the presence of oligonucleotide repeats in promoter regions[29]. Homopolymeric tracts of A or T in potential promoter regions of eighteen genes were found, including eight members of the OMP family (Table 3c).

Virulence

The virulence of individual *H. pylori* isolates has been measured by their ability to produce a cytotoxin-associated protein (CagA) and an active vacuolating cytotoxin (VacA)[5]. The *cagA* gene, though not a virulence determinant, is positioned at one end of a pathogenecity island containing genes that elicit the production of interleukin (IL)-8 by gastric epithelial cells[11,30]. Consistent with its more virulent character, *H. pylori* strain 26695 contains a single contiguous PAI region[11] (Fig. 4).

VacA induces the formation of acidic vacuoles in host epithelial cells, and its presence is associated epidemiologically with tissue damage and disease[31]. VacA may not be the only ulcer-causing factor as 40% of *H. pylori* strains do not produce detectable amounts of the cytotoxin *in vitro*[5]. Sequence differences at the amino terminus and central sections are noted among VacA proteins derived from Tox^+ and Tox^- strains[31]. This Tox^+ *H. pylori* strain contains the more toxigenic S1a/m1 type cytotoxin and three additional large proteins with moderate similarities to the carboxy-terminal end of the active

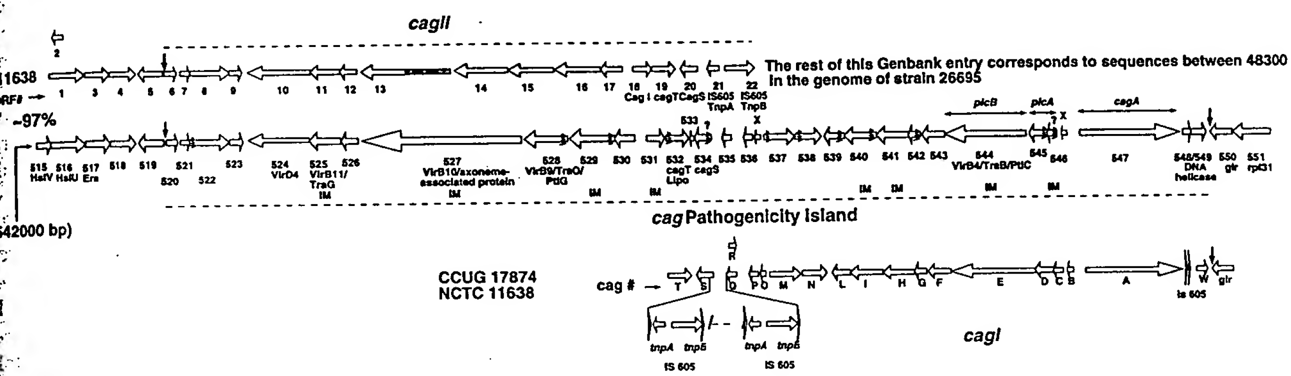


Figure 4 Comparison between the Cag pathogenicity islands of the sequenced strain 26695 and the NCTC11638 strain. The twenty nine ORFs of the contiguous PAI in strain 26695 are represented together with the corresponding ORFs from the PAI present in NCTC11638 (AC000108 and U60176). The PAI in NCTC11638 is divided by the IS 605 elements into two regions, *cagI* and *cagII*. The PAI in NCTC11638 is flanked by a 31-bp (TTACAATTTGAGCCCATTCTTTAGCTTGTTTT) direct repeat (vertical arrows) as described[11]. Some of the genes encode proteins with similarity to proteins involved either in DNA transfer (Vir and Tra proteins) or in export of a toxin (Ptl protein)[10]. However, these genes do not have the conserved contiguous arrangement found in the VirB, Tra and Ptl operons, suggesting that this PAI is not derived from these systems. Most genes of the PAI have no database match, contrary to a previous suggestion[11]. Thirteen of the proteins have a signal peptide (squiggle line), three of them with a weaker probability (squiggled line?). The average length of the signal peptides is 25 amino acids, suggesting that this PAI is of Gram-negative origin. Eight proteins are predicted to have at least two membrane-spanning domains and to be integral membrane proteins (IM)[47]. Although the two PAI are ~97% identical at the nucleotide level, there are several notable and perhaps biologically relevant differences between the two sequences. Four of the genes differ in size. In the PAI of strain 26695, HP 520 and 521 are shorter, whereas HP523 is longer, and HP 527 actually spans both ORF13 and 14. In addition, the N-terminal part of HP527 is 129 amino acids longer than the corresponding region in ORF14. HP548/549 contains a frameshift and is therefore probably inactive in strain 26695. The stippled box preceding ORF13 represents an N-terminal extension not annotated in the Genbank entry for the PAI of NCTC11638. The 'x' indicates ORFs that are neither GeneMark-positive nor GeneSmith-positive, so were not included in our gene list. However, these ORFs may be biologically significant. We do not represent cagR as an ORF, because it is completely contained within ORFQ, and is GeneMark-negative.

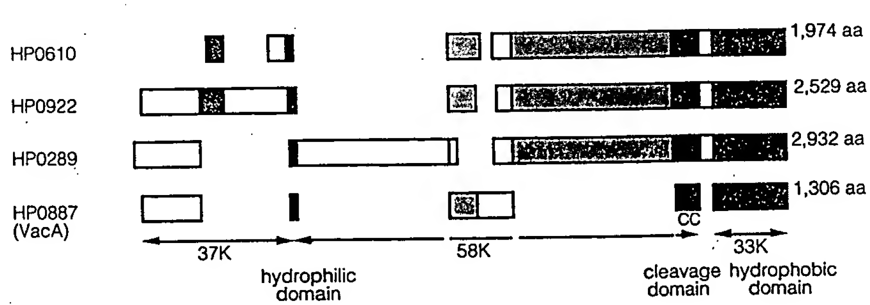


Figure 5 Conserved domains of VacA and related proteins. HP887 is the vacuolating cytotoxin (vacA) gene from *H. pylori* 26695 strain. HP610, HP922 and HP289 are related proteins. Blocks of aligned sequence and the length of each protein are shown. Arrows designate the extents of each VacA domain. The hydrophilic domain (blue boxes) contains the site in VacA at which the N-terminal domain is cleaved into 37K and 58K fragments. The putative cleavage site (ANNNQQNS) differs from that of three cytotoxic strains (CCUG 1784, 60190, G39; AKNDKXES) and is not conserved in the other three VacA-related proteins. The cleavage domain (black boxes) of VacA contains a pair of Cys residues 60 residues upstream from the site at which the C terminus is cleaved. These residues are not conserved in the other three proteins. The 33K C-terminal hydrophobic domain (red boxes) in VacA is thought to form a pore through which the toxin is secreted. The other three proteins show 26-31% sequence similarity to VacA in this region. The other coloured boxes represent regions of similarity.

cytotoxin (~26–31%) (Fig. 5). However, they lack the paired-cysteine residues and the cleavage site required for release of the VacA toxin from the bacterial membrane[31] (Fig. 5). We propose that these proteins may be retained on the outside surface of the cell membrane and contribute to the interaction between *H. pylori* and host cells.

The surface-exposed lipopolysaccharide (LPS) molecule plays an important role in *H. pylori* pathogenesis[32]. The LPS of *H. pylori* is several orders of magnitude less immunogenic than that of enteric bacteria[33] and the O antigen of many *H. pylori* isolates is known to mimic the human Lewis^x and Lewis^y blood group antigen[32]. Genes for synthesis of the lipid A molecule, the core region, and the O antigen were identified. Two genes with low similarity to fucosyltransferases (HP379, HP651) were found and may play a role in the LPS-Lewis antigen molecular mimicry. Our analysis also suggests that three genes, two glycosyltransferases (HP208 and HP619) and one fucosyltransferase (HP379), may be subject to phase variation (Table 3a, b).

As with other pathogens, *H. pylori* probably requires an iron-scavenging system for survival in the host[5]. Genome analysis suggests that *H. pylori* has several systems for iron uptake. One is analogous to the siderophore-mediated iron-uptake *fec* system of *E. coli*[34], except that it lacks the two regulatory proteins (FecR and FecI) and is not organized in a single operon. Unlike other studied systems, *H. pylori* has three copies of each of *fecA*, *exbB* and *exbD*. A second system, consisting of a *feoB*-like gene without *feoA*, suggests that *H. pylori* can assimilate ferrous iron in a fashion similar to the anaerobic feo system of *E. coli*. Other systems for iron uptake present in *H. pylori* consist of the three *frpB* genes which encode proteins similar to either haem- or lactoferrin-binding proteins. Finally, *H. pylori* contains NapA, a bacterioferritin[34], and Pfr, a non-haem cytoplasmic iron-containing ferritin used for storage of iron[35]. The global ferric uptake regulator (Fur) characterized in other bacteria is also present in *H. pylori*. Consensus sequences for Fur-binding boxes were found upstream of two *f*[...] genes, the three *frnB* genes and *fur*.

H. pylori motility is essential for colonization[36]. It enables [...] bacterium to spread into the viscous mucous layer covering [...] gastric epithelium. At least forty proteins in the *H. pylori* geno[...] appear to be involved in the regulation, secretion and assembly [...] the flagellar architecture. As has bene reported for the *flaA* and *fl*[...] genes, we identified sigma 28 and sigma 54-like promoter eleme[...] upstream of many flagellar genes, underscoring the complexity [...] the transcriptional regulation of the flagellar regulon[5].

Acidity, pH and acid tolerance

H. pylori is unusual among pathogenic bacteria in its ability [...] colonize host cells in an environment of high acidity. As it enters [...] gastric environment by oral ingestion, the organism is transien[...] subjected to the extreme pH of the lumen side of the gastric muc[...] layer (pH ~2). The survival of *H. pylori* in acidic environments [...] probably due to its ability to establish a positive inside-membra[...] potential[37] and subsequently to modify its microenvironme[...] through the action of urease and the release of factors that inhi[...] acid production by parietal cells[5]. A switch in membrane polar[...] provides an electrical barrier that prevents the entry of proto[...] (H^+). A positive cell interior can be created by the active extrusion [...] anions or by a proton diffusion potential. The latter model appe[...] more likely as no clear mechanism for electrogenic anion efflux [...] apparent in the genome. A proton diffusion potential would requ[...] the anion permeability of the cytoplasmic membrane to be low an[...] thus far, only three anion transporters have been identified. How[...]ever, it remains to be determined whether anion conductances [...] associated with other proteins: the MDR-like transporters (HP60[...] HP1082 and HP1206) or hypotheticals. Although it has be[...] suggested that proton-translocating P-type ATPases could media[...] survival in acid conditions by the extrusion of protons from t[...] cytoplasm[38], this idea is not supported by the identified transport[...]

Table 3 Homopolymeric tracts and dinucleotide repeats in *H. pylori*

HP no.	ID	No. of repeats	Gene status	Poly(A) or Poly(T) tracts in 5′ intergenic reg[...]
9	OMP	11 CT	Off	Poly(A)
208	glycos. transf.	11 AG	Truncated	Poly(A)
638	OMP	6 CT	On	No
722	OMP	8 CT	Off	Poly(T)
725	OMP	6 CT	Off	Poly(T)
744	Hypo	9 AG	Truncated	No
896	OMP	11 CT	On	Poly(A)
1417	Cons. Hypo	9 AG	Truncated	No

Nucleotide sequence at the beginning of HP0722 showing the CT dinucleotide repeat and the poly T tract. The putative ribosome binding site is shown in green. Translat[...] starting at the designated methionine leads to a truncated product. The addition or deletion of two CT repeats, by 'slipped-strand mispairing', will restore the frame.

```
CCAAAAATCTTTTTTTTTTTTTTTTTTGAAATCCAATAAATTTATGGTAAAGT-37bp-TTTACAATAAAAAAATTACTTTAAGGAACATTT
TATGAAAAAGACAATTCTACTCTCTCTCTCTCTCTCGCTTCATCGCTCTTGCACGCTGAAGACAACGGCTTTTTTGTGAGCGCCGGCT
Y  E  K  D  N  S  T  L  S  L  S  L  A  S  S  L  L  H  A  E  D  N  G  F  F  V  S  A  G  Y
 M  K  K  T  I  L  L  S  L  S  L  S  L  H  R  S  C  T  L  K  T  T  A  F  L  *
```

(b) Homopolymeric poly(C) and poly(G) tracts within coding sequence

HP no.	ID	Tract length	Gene stat[...]
58	Hypo	C15	Off
217	Hypo	G12	On
379	fucosyl transf.	C13	On
464	TypeI R	C15	On
619	glycos. transf.	C13	Truncate[...]
651	Hypo	C13	On
1353	Hypo	C15	Truncate[...]
1471	TypeIIS-R	G14	On
1522	Methyl ase	G12	Truncate[...]

Genes possibly regulated by homopolymeric poly(A) or poly(T) tracts in 5′ intergenic regions

HP no.	ID	Tract	HP no.	ID	Tract	HP no.	ID	Tr[...]
9	OMP	A14	25	OMP	T15	208	*rfaJ*	A[...]
227	OMP	T14	228	IMP	A14	349	*pyrG*	T[...]
350	IMP	A15	547	*cagA*	A14	629	Hypo	T[...]
722	OMP	T16	725	OMP	T14	733	Hypo	T[...]
876	*frpB*	T16	896	OMP	A14	912	OMP	T[...]
1342	OMP	A14	1400	*fecA*	A16			

The P-type ATPase sequences in *H. pylori* (*copAP*, HP791, and HP1503) are more closely related to divalent cation transporters than to ATPases with specificity for protons or monovalent cations. One of them, HP0791, is involved in Ni^{2+} supply, an essential component of urease activity[39]. The others may be involved in the elimination of toxic metals from the cytoplasm and not in pH regulation.

Additional mechanisms of pH homeostasis may well contribute to *H. pylori* survival. A change in protein content observed in response to a shift of extracellular pH from 7.5 to 3.0 suggests the presence of an acid-inducible response[40]. Although *H. pylori* lacks most orthologues of the genes that are acid-induced in *E. coli* and *Salmonella typhimurium*, including the amino-acid decarboxylases and formate hydrogen lyase, certain virulence factors, outer membrane import under anaerobic conditions for cytochromes, catalase).

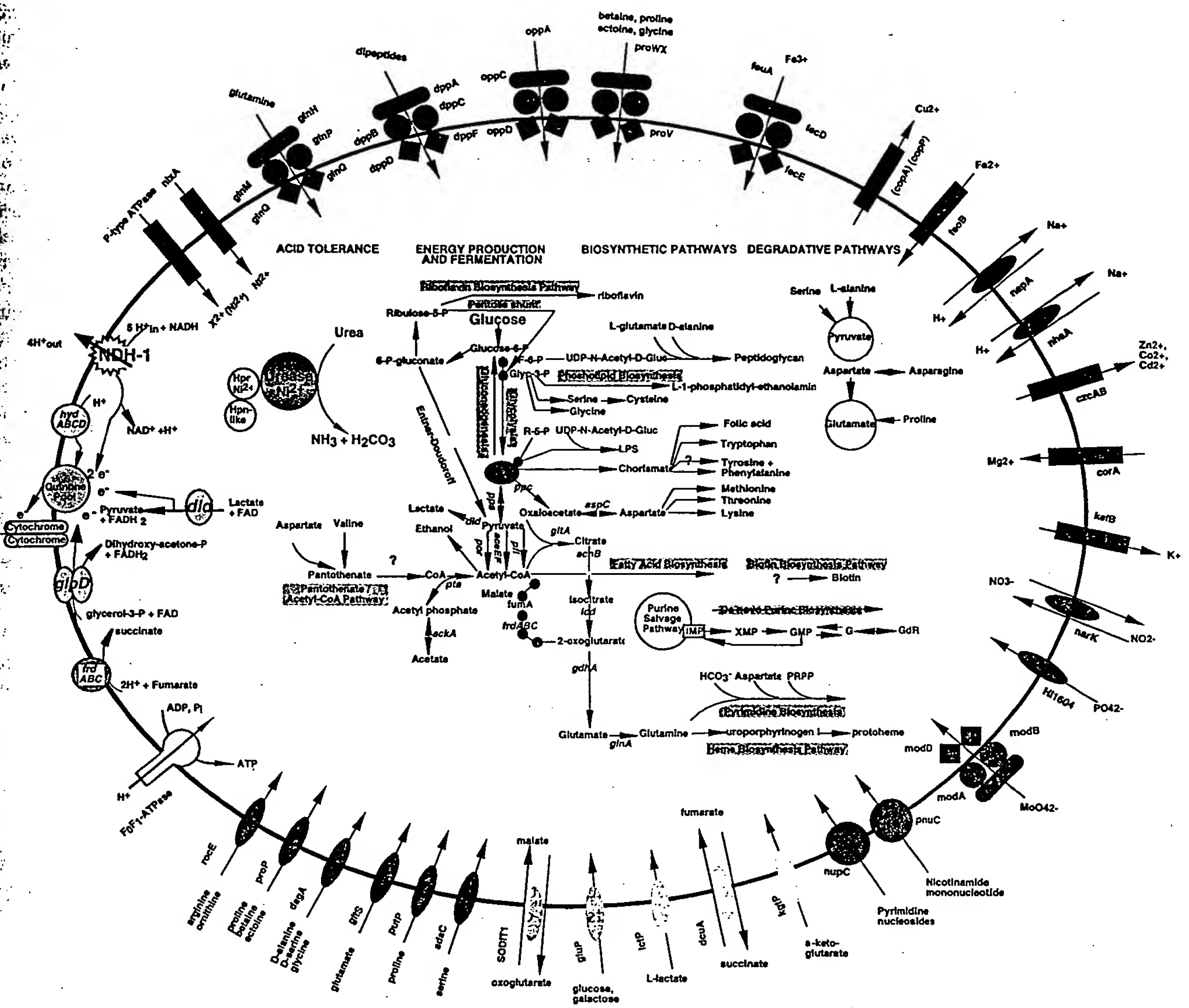


Figure 6 Solute transport and metabolic pathways of *Helicobacter pylori*. Transporters identified by sequence comparisons are characteristc of Gram-negative bacteria. Colours correspond to transport role categories defined by Riley[15]: blue, amino acids, peptides and amines; red, anions; yellow, carbohydrates, organic alcohols and acids; green, cations; and purple, nucleosides, purines and pyrimidines. Numerous permeases (ovals) with specificity for amino acids (*recE*, *proP*, *dagA*, *gltS*, *putP* and *sdaC*) or carbohydrates (*SODiTI*, *gluP*, *lactP*, *cduA*, *kgtP*) import organic nutrients. Structurally related permease proteins maintain ionic homeostasis by transporting HPO_4^{2-} (*HI1604*), NO_3^{2-} (*narK*), and Na^+ (*nhA*, *napA*). Primary active-transport systems, independent of the proton cycle, are also apparent. Included in this group are ATP-binding protein-cassette (ABC) transporters (composite figures of 2 diamonds, 2 circles, 1 oval) for the uptake of oligopeptides (*oppACD*), dipeptides (*dppABCDF*), proline (*proVWX*), glutamine (*glnHMPQ*), molybdenum (*modABD*), and iron III (*fecED*), P-type ATPases that extrude toxic metals from the cell (*copAP* and *cadA*), and the glutathione-regulated potassium-efflux protein (*kefB*). Transporters for the accumulation of ionic cofactors are encoded by *nixA* (Ni^{2+} for urease activation), *corA* (Mg^{2+} for phosphohydrolases, phosphotransferases, ATPases) and *feoB* (Fe^{2+} import under anaerobic conditions for cytochromes, catalase). An integrated view of the main components of the central metabolism of *H. pylori* strain 26695 is presented. The use of glucose as the sole carbohydrate source is emphasized. Urease, a multisubunit Ni^{2+}-binding enzyme, is crucial for colonization and for survival of *H. pylori* at acid pH, and is indicated as a complex (purple circle) with Hpn, a Ni^{2+}-binding cofactor, and a newly identified Hpn-like protein (HP1432). A question mark is attached to pathways that could not be completely elucidated. Pathways or steps for which no enzymes were identified are represented by a red arrow. Pathways for macromolecular biosynthesis (RNA, DNA and fatty acids) have been omitted. *ackA*, acetate kinase; *acnB*, aconitase B; *aspC*, aspartate aminotransferase; *dld*, D-lactate dehydrogenase; *gdhA*, glutamate dehydrogenase; *glnA*, glutamine synthetase; *gltA* citrate synthase; *HydABC*, hydrogenase complex; *icd*, isocitrate dehydrogenase; *pfl*, pyruvate formate lyase; *por*, pyruvate ferredoxin oxidoreductase; *ppc*, phosphoenolpyruvate carboxylase; *pps*, phosphoenolpyruvate synthase; *pta*, phosphate acetyltransferase; *gldD*, glycerol-3-phosphate dehydrogenase; NDH-1, NADH-ubiquinone oxidoreductase complex.

proteins, sensor-regulator pairs and other proteins may be acid-induced.

Regulation of gene expression

Bacteria regulate the transcription of their genes in response to many environmental stimuli, such as nutrient availability, cell density, pH, contact with target tissue, DNA-damaging agents, temperature and osmolarity. In the case of pathogens, the regulated expression of certain key genes is essential for successful evasion of host responses and colonization, adaptation to different body sites, and survival as the pathogen passes to new hosts. In *H. pylori*, global regulatory proteins are less abundant than in *E. coli*. For example, orthologues of many DNA-binding proteins that regulate the expression of certain operons such as OxyR (oxidative stress), Crp (carbon utilization), RpoH (heat shock), and Fnr (fumarate and nitrate regulation) are absent. Only four *H. pylori* proteins have a perfect match to helix–turn–helix (HTH) motifs, a signature of transcription factors; a putative heat-shock protein (HspR), two proteins with no database match (HP1124 and HP1349) and SecA, a component of the general secretory machinery. In contrast, 34 proteins containing an HTH motif were found in *H. influenzae* and 148 in *E. coli*. We identified several other putative regulatory functions, including SpoT and CstA for 'stringent response' to amino-acid starvation and to carbon starvation, respectively.

Environmental response requires sensing changes and transmission of this information to cellular regulatory networks. Two-component regulator systems, consisting of a membrane histidine kinase sensor protein and a cytoplasmic DNA-binding response regulator, provide a well studied mechanism for such signal transduction. Four sensor proteins and seven response regulators were found in *H. pylori*, similar to the number found in *H. influenzae*[7]. This is approximately one third the number found in *E. coli* which, in contrast to *H. pylori* and *H. influenzae*, may be exposed to more environments.

Metabolism

Metabolic pathway analysis of the *H. pylori* genome suggests the following features. *H. pylori* uses glucose as the only source of carbohydrate and the main source for substrate-level phosphorylation. It also derives energy from the degradation of serine, alanine, aspartate and proline. The glycolysis–gluconeogenesis metabolic axis constitutes the backbone of energy production and the start point of many biosynthetic pathways. The biosynthesis of peptidoglycan, phospholipids, aromatic amino acids, fatty acids and cofactors is derived from acetyl-CoA or from intermediates in the glycolytic pathway (Fig. 6). The metabolism of pyruvate reflects the microaerophilic character of this organism. Neither the aerobic pyruvate dehydrogenase (*aceEF*) nor the strictly anaerobic pyruvate formate lyase (*pfl*) associated with mixed-acid fermentation are present. The conversion of pyruvate to acetyl CoA is performed by the pyruvate ferrodoxin oxidoreductase (POR), a four-subunit enzyme thus far only described in hyperthermophilic organisms[41]. The tricarboxylic acid cycle (TCA) is incomplete and the glyoxylate shunt is absent. The analysis of degradative pathways, uptake systems and biosynthetic pathways for pyrimidine, purine and haem suggests that *H. pylori* uses several substrates as nitrogen source, including urea, ammonia, alanine, serine and glutamine. The assimilation of ammonia, an abundant product of urease activity, is achieved by the glutamine synthase enzyme and α-ketoglutarate is transformed into glutamate by glutamate dehydrogenase rather than by the glutamate synthase enzyme.

In *H. pylori*, proton translocation is mediated by the NDH-1 dehydrogenase and the different cytochromes, including the primitive-type cytochrome cbb3 (Table 2). Four respiratory electron-generating deydrogenases have been identified, glycerol-3-phosphate dehydrogenase (GlpD), D-lactate dehydrogenase, NADH–ubiquinone oxidoreductase complex (NDH-1), and a hydrogenase complex (HydABC). Our analysis also suggests that *H. pylori* is not able to use nitrate, nitrite, dimethylsulphoxide, trimethylamine *N*-oxide or thiosulphate as electron acceptors. Much of our metabolic analysis is supported by experimental evidence[41,42].

Evolutionary relationships of *H. pylori*

H. pylori is currently classified in the Proteobacteria, a large, diverse division of Gram-negative bacteria which includes two other completely sequenced species, *H. influenzae* and *E. coli*. Given this taxonomic placement, based primarily on 16S rRNA sequence comparisons, one might expect the proteins of *H. pylori* most closely to resemble their *H. influenzae* and *E. coli* homologues rather than those in other genomes such as *Synechocystis* sp., *M. genitalium*, *M. pneumoniae*, *M. jannaschii*, and *Saccharomyces cerevisae*. This is indeed the case for many proteins. There are, however, many examples of *H. pylori* proteins in amino-acid biosynthesis, energy metabolism, translation and cellular processes that have greater sequence similarity to those found in non-Proteobacteria. For example, Dhs1, the initial enzyme in the chorismate biosynthesis pathway is 75.5% similar to *Arabidopsis thaliana* chloroplast Dhs1 gene product, and has minimal sequence similarity to the equivalent *E. coli* AroH, AroF or AroG gene products. The remaining enzymes in this pathway have strong sequence similarity to their *E. coli* counterpart. Similarly, the *H. pylori* prephenate dehydrogenase (TyrA), which converts chorismate to tyrosine, and six out of 15 enzymes in the aspartate amino-acid biosynthetic pathways, resemble those from *B. subtilis*. A similar pattern can be seen in a different functional category. Nearly all *H. pylori* tRNA synthetases have eubacterial homologues, mostly with best matches to Proteobacteria species. However, histidyl-tRNA synthetase shows several amino-acid sequence signatures in common with eukaryotic and archaeal (*M. jannaschii*) homologues.

Such observations of discordant sequence similarity are often interpreted as evidence of lateral gene transfer in the evolutionary history of an organism. It is also possible that *H. pylori* diverged early from the lineage that led to the gamma Proteobacteria, and retained more ancient forms of enzymes that have been subsequently replaced or have diverged extensively in *H. influenzae* and *E. coli*.

Conclusion

Our whole-genome analysis of *H. pylori* gives new insight into its pathogenesis, acid tolerance, antigenic variation and microaerophilic character. The availability of the complete genome sequence will allow further assessment of *H. pylori* genetic diversity. This is an important aspect of *H. pylori* epidemiology as allelic polymorphism within several loci has already been associated with disease outcome[5,21,31]. The extent of molecular mimicry between *H. pylori* and its human host, an underappreciated topic, can now be fully explored[43]. The identification of many new putative virulence determinants should allow critical tests of their roles and thus new insight into mechanisms of initial colonization, persistence of this bacterium during long-term carriage, and the mechanisms by which it promotes various gastroduodenal diseases.

Methods

H. pylori strain 26695 (ref. 44) was originally isolated from a patient in the United Kingdom with gastritis (K. Eaton, personal communication) and was chosen because it colonizes piglets and elicits immune and inflammatory responses. It is also toxigenic, and transformable, and thus amenable to mutational tests of gene function.

The *H. pylori* genome sequence was obtained by a whole-genome random sequencing method previously applied to genomes of *Haemophilus influenzae*, *Mycoplasma genitalium*[8], and *Methanococcus jannaschii*[9]. Ninety-two per cent of the genome was covered by at least one λ clone and only 0.56% of the genome had single-fold coverage.

Open reading frames (ORFs) and predicted coding regions were identified using three methods. The predicted protein-coding regions were initially defined by searching for ORFs longer than 80 codons. Coding potential analysis of the entire genome was performed with a version of GeneMark[45] trained with a set of *H. pylori* ORFs longer than 600 nucleotides. Coding sequences and potential starts of translation were also determined using GeneSmith (H.S., unpublished), a program that evaluates ORF length, separation of ORFs and overlap and quality of ribosome binding site. ORFs with low GeneMark coding potential, no database match, and not retained by GeneSmith were eliminated. GeneSmith identified 25 ORFs that are smaller than 100 codons, had no database match and were GeneMark negative. Frameshifts were detected by inspecting pairwise alignments, families of orthologues (similar proteins derived from different species) and paralogues (similar proteins from within the same organism), and regions containing homopolymer stretches and dinucleotide repeats. Ambiguities were resolved by an alternative sequencing chemistry (terminator reactions), and by sequencing PCR products obtained using the genomic DNA as template. Frameshifts that remain in the genome are considered authentic and not sequencing artefacts.

To determine their identity, ORFs were searched against a non-redundant amino-acid database as previously described[9]. ORFs were also analysed using 175 hidden Markov models constructed for a number of conserved protein families (pfam v1.0) using hmmer[43]. In addition, all ORFs were searched against the prosite motif database using MacPattern[46]. Families of paralogues were constructed by pairwise searches of proteins using FASTA. Matches that spanned at least 60% of the smaller of the protein pair were retained and visually inspected.

A unix version of the program TopPred[47] was used to identify membrane-spanning domains (MSD) in proteins. Six hundred and sixty three proteins containing at least one MSD were found; of these, 300 had 2 potential MSDs or more. The presence of signal peptides and the probable position of the cleavage site in secreted proteins were detected using Signal-P, a neural net program that had been trained on a curated set of secreted proteins from Gram-negative bacteria[48]. 367 proteins were predicted to have a signal peptide. Lipoproteins were identified by scanning for the presence of a lipobox in the first 30 amino acids of every protein; 20 lipoproteins were identified, eighteen of which were Signal-P positive. Outer-membrane proteins were found by searching for aromatic amino acids at the end of the proteins.

Homopolymer and dinucleotide repeats were found by using RepScan (H.O.S., unpublished) which finds direct repeats of any length. All features identified using these programs were validated by visual inspection to remove false positives. Metabolic pathways were curated by hand and by reference to EcoCyc[49].

Received 16 May; accepted 1 July 1997.

1. Warren, J. R. & Marshall, B. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet* 1, 1273–1275 (1983).
2. Cover, T. L. & Blaser, M. J. *Helicobacter pylori* infection, a paradigm for chronic mucosal inflammation: pathogenesis and implications for eradication and prevention. *Adv. Int. Med.* 41, 85–117 (1996).
3. Mobley, H. L. T., Island, M. D. & Hausinger, R. P. Molecular Biology of Microbial Ureases. *Microbiol. Rev.* 59, 451–480 (1995).
4. Go, M. F. & Graham, D. Y. How dos *Helicobacter pylori* cause duodenal ulcer disease: The bug, the host, or both? *J. Gastroentrol. Hepatol.* (suppl.) 9, 8–12 (1994).
5. Labigne, A. & de Reuse, H. Determinants of *Helicobacter pylori* pathogenicity. *Infect. Agents Disease* 5, 191–202 (1996).
6. Clemens, J. *et al.* Impact of infection by *Helicobacter pylori* on the risk and severity of endemic cholera. *J. Inf. Dis.* 171, 1653–1656 (1995).
7. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512 (1995).
8. Fraser, C. M. *et al.* The *Mycoplasma genitalium* genome sequence reveals a minimal gene complement. *Science* 270, 397–403 (1995).
9. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073 (1996).
10. Winans, S. C., Burns, D. L. & Christie, P. J. Adaptation of a conjugal transfer system for the export of pathogenic macromolecules. *Trends Microbiol.* 4, 64–68 (1996).
11. Censini, S. *et al.* Cag, a pathogenicity island of *Helicobacter pylori*, encodes typeI-specific and disease-associated virulence factors. *Proc. Natl Acad. Sci. USA* 93, 14648–14653 (1996).
12. http://genome.wustl.edu/eddy/low/tRNAscan-SE-Manual/Manual.html
13. Akopyants, N. S., Kersulyte, D. & Berg, D. E. DNA rearrangement in the 40 kb cag (virulence) region in the *Helicobacter pylori* genome. *Gut* 39 (suppl. 2), A67 (1996).
14. Marczynski, G. T. & Shapiro, L. Bacterial chromosome origins of replication. *Curr. Opin. Gen. Dev.* 3, 775–782 (1993).
15. Riley, M. Functions of gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952 (1993).
16. Kornberg, A. & Baker, T. A. Replication mechanisms and operations in DNA replication. (ed. Kornberg, A. & Baker, T.) 471–510 (Freeman, New York, 1992).
17. Macnab, R. M. in Escherichia coli *and* Salmonella *Cellular and Molecular Biology* (eds Neidhardt, F. C. *et al.*) 123–145 (ASM, Washington DC, 1996).
18. Strom, M. S., Nunn, D. N. & Lory, S. Posttranslational processing of type IV prepilin and homologs by PilD of Pseudomonas aeruginosa. *Meth. Enzymol.* 235, 527–540 (1994).
19. Bardwell, J. C. Building bridges: disulphide bond formation in the cell. *Mol. Microbiol.* 14, 199–205 (1994).
20. Linn, S. in Escherichia coli and *Salmonella Cellular and Molecular Biology* (eds Neidhardt, F. C. *et al.*) 764–772 (ASM, Washington D.C., 1996).
21. Peek, R. M., Thompson, S. A., Atherton, J. C., Blaser, M. J. & Miller, G. G. Expression of iceA, a novel ulcer-associated *Helicobacter pylori* gene, is induced by contact with gastric epithelial cells and is associated with enhanced mucosal IL-8. *Gut* 39 (suppl. 2), A71 (1996).
22. Curnow, A. W., Ibba, M. & Soll, D. tRNA-dependent asparagine formation. *Nature* 382, 589–590 (1996).
23. Jones, A. C., Foynes, S., Cockayne, A. & Penn, C. W. Gene cloning of a flagellar sheath protein of *Helicobacter pylori* shows its identity with the putative adhesin, HpaA. *Gut* 39 (suppl. 2), A62 (1996).
24. Boren, T., Falk, P., Roth, K. A., Larson, G. & Normark, S. Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. *Science* 262, 1892–1895 (1993).
25. Ilver, D. *et al.* The *Helicobacter pylori* blood group antigen binding adhesin. *Gut* 39 (suppl. 2), A55 (1996).
26. Odenbreit, S., Till, M. & Haas, R. Optimized blaM-transposon shuttle mutagenesis of *Helicobacter pylori* allows identification of novel genetic loci involved in bacterial virulence. *Mol. Microbiol.* 20, 361–373 (1996).
27. Peterson, S. N. *et al.* Characterization of repetitive DNA in the *Mycoplasma genitalium* genome: possible role in the generation of antigenic variation. *Proc. Natl Acad. Sci. USA* 92, 11829–11833 (1995).
28. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* 4, 24–33 (1994).
29. Jonsson, A. B., Nyberg, G. & Normark, S. Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. *EMBO J.* 10, 477–488 (1991).
30. Tummuru, M. K. R., Sharma, S. A. & Blaser, M. J. *Helicobacter pylori* picB, a homologue of the *Bordetella pertussis* toxin secretion protein, is required for induction of IL-8 in gastric epithelial cells. *Mol. Microbiol.* 18, 867–876 (1995).
31. Atherton, J. C. *et al.* Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. Association of specific vacA types with cytotoxin production and peptic ulceration. *J. Biol. Chem.* 270, 17771–17777 (1995).
32. Moran, A. P. The role of lipopolysaccharide in *Helicobacter pylori* pathogenesis. *Aliment. Pharmacol. Ther.* 10 (suppl. 1), 39–50 (1996).
33. Baker, P. J. *et al.* Molecular structures that influence the immunomodulatory properties of the lipid A and inner core region oligosaccharides of bacterial lipopolysaccharides. *Infect. Immun.* 62, 2257–2269 (1994).
34. Earhart, C. F. in Escherichia coli *and* Salmonella *Cellular and Molecular Biology* (eds Neidhardt, F. C. *et al.*) 1075–1090 (ASM, Washington DC, 1996).
35. Evans, D. J. Jr, Evans, D. G., Lampert, H. C. & Nakano, H. Identification of four new prokaryotic bacterioferritins, from *Helicobacter pylori*, *Anabaena variabilism*, *Bacillus subtilis* and *Treponema pallidum*, by analysis of gene sequences. *Gene* 153, 123–127 (1995); Frazier, B. A. *et al.* Paracrystalline inclusions of a novel ferritin containing nonheme iron, produced by the human gastric pathogen *Helicobacter pylori*: evidence for a third class of ferritins. *J. Bacteriol.* 175, 966–972 (1993).
36. Suerbaum, S. The complex flagella of gastric *Helicobacter* species. *Trends Microbiol.* 3, 168–170 (1995).
37. Matin, A., Zychlinsky, E., Keyhan, M. & Sachs, G. Capacity of *Helicobacter pylori* to generate ionic gradients at low pH is similar to that of bacteria which grow under strongly acidic conditions. *Infect. Immun.* 64, 1434–1436 (1996).
38. Melchers, K. *et al.* Cloning and membrane topology of a P type ATPase from *Helicobacter pyroli*. *J. Biol. Chem.* 271, 446–457 (1996).
39. Melchers, K. *et al.* Cloning and analysis of two P type ion pumps of *Helicobacter pylori*, a cation resistance ATPase and a membrane pump necessary for urease activity. *Gut* 39 (suppl. 2), A67 (1996).
40. McGowan, C. C., Cover, T. L. & Blaser, M. J. *Helicobacter pylori* and gastric acid: biological and therapeutic implications. *Gastroenterology* 110, 926–938 (1996).
41. Hughes, N. J., Chalk, T. L., Clayton, C. L. & Kelly, D. J. Identification of carboxylation enzymes and characterization of a novel four-subunit pyruvate:flavodoxin oxidoreductase from *Helicobacter pylori*. *J. Bacteriol.* 177, 3953–3959 (1995).
42. Mendz, G. L. & Hazell, S. L. Aminoacid utilizaiton by *Helicobacter pylori*. *Int. J. Biochem. Cell. Biol.* 27, 1085–1093 (1995).
43. Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. Pfam: A comprehensive database of protein families based on seed alignments. *Proteins* (in the press).
44. Akopyants, N. S., Eaton, K. A. & Berg, D. E. Adaptive mutation and co-colonization during *Helicobacter pylori* infection of gnotobiotic piglets. *Infect. Immun.* 63, 116–121 (1995).
45. Borodovsky, M., Rudd, K. E. & Koonin, E. V. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.* 22, 4756–4767 (1994).
46. Fuchs, R. MacPattern: protein pattern searching on the Apple MacIntosh. *Comput. Appl. Biosci.* 7, 105–106 (1991).
47. Claros, M. G. & von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* 10, 685–686 (1994).
48. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6 (1997).
49. Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* 25, 43–51 (1997).
50. Doig, P., Exner, M. M., Hancock, R. E. & Trust, T. J. Isolation and characterization of a conserved porin protein from *Helicobacter pylori*. *J. Bacteriol.* 177, 5447–5452 (1995).

Acknowledgements. D.E.B., M.B. and W.H. are supported by grants from the NIH; P.K. is supported by a grant from the National Center for Research Resources. We thank N. S. Akopyants for preparing high quality chromosomal DNA from *H. pylori* strain 26695; M. Heaney, J. Scott, A. Saeed and R. Shirley for software and database support; and V. Sapiro, B. Vincent, J. Meehan and D. Mass for computer system support.

Correspondence and requests for materials should be addressed to J.-F.T. (e-mail: ghp@tigr.org). The annotated genome sequence and gene family alignments are available on the World-Wide Web site at http://www.tigr.org/tdb/mdb/hpdb/hpdb.html. The sequence has been deposited with GenBank under accession number AE000511.

Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*

Claire M. Fraser*, Sherwood Casjens†, Wai Mun Huang†, Granger G. Sutton*, Rebecca Clayton*, Raju Lathigra‡, Owen White*, Karen A. Ketchum*, Robert Dodson*, Erin K. Hickey*, Michelle Gwinn*, Brian Dougherty*, Jean-Francois Tomb*, Robert D. Fleischmann*, Delwood Richardson*, Jeremy Peterson*, Anthony R. Kerlavage*, John Quackenbush*, Steven Salzberg*, Mark Hanson‡, Rene van Vugt†, Nanette Palmer†, Mark D. Adams*, Jeannine Gocayne*, Janice Weidman*, Teresa Utterback*, Larry Watthey*, Lisa McDonald*, Patricia Artiach*, Cheryl Bowman*, Stacey Garland*, Claire Fujii*, Matthew D. Cotton*, Kurt Horst*, Kevin Roberts*, Bonnie Hatch*, Hamilton O. Smith* & J. Craig Venter*

* *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
† *Division of Molecular Biology and Genetics, Department of Oncological Sciences, University of Utah, Salt Lake City, Utah 84132, USA*
‡ *MedImmune, Inc., 35 West Watkins Mill Road, Gaithersburg, Maryland 20878, USA*

The genome of the bacterium *Borrelia burgdorferi* B31, the aetiologic agent of Lyme disease, contains a linear chromosome of 910,725 base pairs and at least 17 linear and circular plasmids with a combined size of more than 533,000 base pairs. The chromosome contains 853 genes encoding a basic set of proteins for DNA replication, transcription, translation, solute transport and energy metabolism, but, like *Mycoplasma genitalium*, it contains no genes for cellular biosynthetic reactions. Because *B. burgdorferi* and *M. genitalium* are distantly related eubacteria, we suggest that their limited metabolic capacities reflect convergent evolution by gene loss from more metabolically competent progenitors. Of 430 genes on 11 plasmids, most have no known biological function; 39% of plasmid genes are paralogues that form 47 gene families. The biological significance of the multiple plasmid-encoded genes is not clear, although they may be involved in antigenic variation or immune evasion.

In the mid-1970s, a geographic clustering of an unusual rheumatoid arthritis-like condition was reported in Connecticut[1]. That cluster of cases focused attention on the syndrome that is now called Lyme disease. It was subsequently realized that a similar disorder had been known in Europe since the beginning of this century. Lyme disease is characterized by some or all of the following manifestations: an initial erythematous annular rash, 'flu-like symptoms, neurological complications, and arthritis in about 50% of untreated patients[2]. In the United States, the disease occurs primarily in northeastern and midwestern states, and in western parts of California and Oregon. These regions coincide with the ranges of various species of *Ixodes* ticks, the primary vector of Lyme disease. Lyme disease is now the most common tick-transmitted illness in the United States, and has been reported in many temperate parts of the Northern Hemisphere.

It was not until the early 1980s that a new spirochaete, *Borrelia burgdorferi*[3], was isolated and cultured from the midgut of *Ixodes* ticks, and subsequently from patients with Lyme disease[4,5]. Analysis of genetic diversity among individual *Borrelia* isolates has defined a closely related cluster containing at least 10 tick-borne species of Lyme disease agents, called '*B. burgdorferi* (*sensu lato*)'. *B. burgdorferi* resembles most other spirochaetes in that it is a highly specialized, motile, two-membrane, spiral-shaped bacterium that lives primarily as an extracellular pathogen. *Borrelia* is fastidious and difficult to culture *in vitro*, requiring a specially enriched media and low oxygen tension[6].

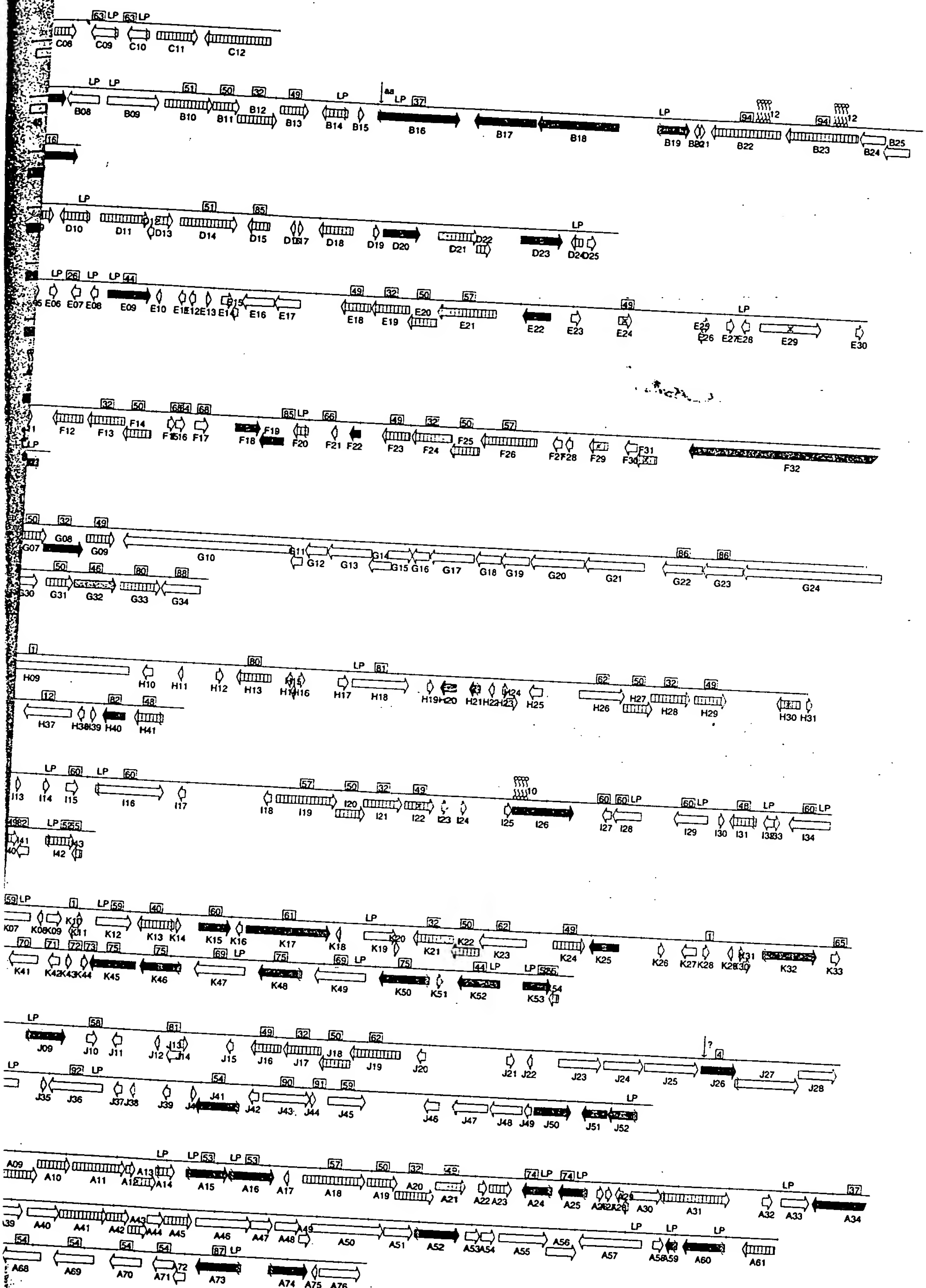
One of the most striking features of *B. burgdorferi* is its unusual genome, which includes a linear chromosome approximately one megabase in size[7–10] and numerous linear and circular plasmids[11–13], with some isolates containing up to 20 different plasmids. The plasmids have a copy number of approximately one per chromosome[10,14], and different plasmids often appear to share regions of homologous DNA[13,15,16]. Long-term culture of *B. burgdorferi* results in the loss of some plasmids, changes in protein expression profiles, and a loss in the ability of the organism to infect laboratory animals, suggesting that the plasmids encode important proteins involved in virulence[17–19].

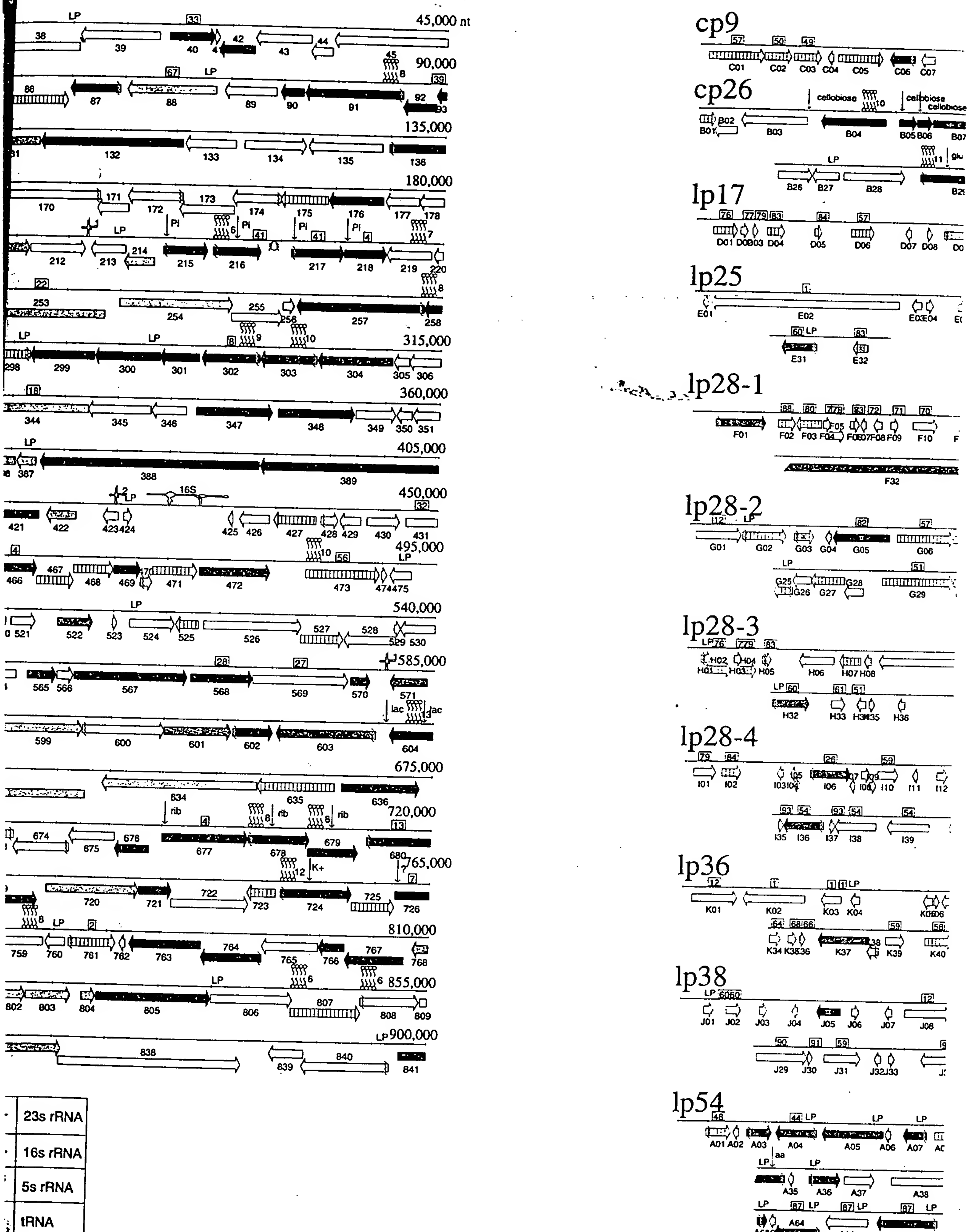
Because of its importance as a pathogen of humans and animals, and the value of complete genome sequence information for understanding its life cycle and advancing drug and vaccine development, we sequenced the genome of *B. burgdorferi* type strain (B31), using the random sequencing method previously described[20–24]. Here we summarize the results from sequencing, assembly and analysis of the linear chromosome and 11 plasmids.

Chromosome analysis

The linear chromosome of *B. burgdorferi* has 910,725 base pairs (bp) and an average G+C content of 28.6%. Base pair one represents the first double-stranded base pair that we observed at the left telomere. Previous genome characterizations agree with the nucleotide sequence of the large chromosome[10,25–28]. The 853 predicted coding sequences (open reading frames; ORFs) have an average size of 992 bp, similar to that observed in other prokaryotic genomes, with 93% of the *B. burgdorferi* genome representing

Figure 1 Linear representations of the *B. burgdorferi* B31 chromosome and plasmids. The location of predicted coding regions colour-coded by biological role, RNA genes, and tRNAs is indicated. Arrows represent the direction of transcription for each predicted coding region. Numbers associated with tRNA symbols represent the number of tRNAs at a locus. Numbers associated with GES represent the number of membrane-spanning domains according to the Goldman, Engelman and Steitz scale as calculated by TopPred[49]. Only proteins with five or more GES are indicated. Members of paralogous gene families are identified by family number. Transporter abbreviations: mal, maltose; P, gly and bet, proline, glycine, betaine; glyc, glycerol; aa, amino acid; E, glutamate; fru, fructose; glu, glucose; s/p, spermidine/putrescine; pan, pantothenate; Pi, phosphate; lac, lactate; rib, ribose; ?, unknown.


63 LP 63 LP
C08 C09 C10 C11 C12
LP LP 51 50 32 49 LP aa LP 37
B08 B09 B10 B11 B12 B13 B14 B15 B16 B17 B18
LP 94 12 94 12
B19 B20 B21 B22 B23 B24 B25
16
LP 51 85
D10 D11 D12 D13 D14 D15 D16 D17 D18 D19 D20 D21 D22 D23 D24 D25
LP 26 LP LP 44
E06 E07 E08 E09 E10 E11 E12 E13 E14 E15 E16 E17
49 32 50 57 49 LP
E18 E19 E20 E21 E22 E23 E24 E25 E26 E27 E28 E29 E30
32 50 68 64 68 85 LP 66 49 32 50 57
F12 F13 F14 F15 F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30 F31 F32
50 32 49
G07 G08 G09 G10 G11 G12 G13 G14 G15 G16 G17 G18 G19 G20 G21
86 86
G22 G23 G24
50 46 80 88
G30 G31 G32 G33 G34
1 80 LP 81
H09 H10 H11 H12 H13 H14 H15 H16 H17 H18 H19 H20 H21 H22 H23 H24 H25
62 50 32 49
H26 H27 H28 H29 H30 H31
12 82 48
H37 H38 H39 H40 H41
LP 60 LP 60
I13 I14 I15 I16 I17
57 50 32 49 10
I18 I19 I20 I21 I22 I23 I24 I25 I26
60 60 LP 60 LP 48 LP 60 LP
I27 I28 I29 I30 I31 I32 I33 I34
49 82 LP 55 55
I40 I41 I42 I43
59 LP 1 LP 59 40 60 61 LP 32 50 62 49 1 65
K07 K08 K09 K10 K11 K12 K13 K14 K15 K16 K17 K18 K19 K20 K21 K22 K23 K24 K25 K26 K27 K28 K29 K30 K31 K32 K33
70 71 72 73 75 75 69 LP 75 69 LP 75 44 LP LP 55 55
K41 K42 K43 K44 K45 K46 K47 K48 K49 K50 K51 K52 K53 K54
LP 58 81 49 32 50 62 ? 4
J09 J10 J11 J12 J13 J14 J15 J16 J17 J18 J19 J20 J21 J22 J23 J24 J25 J26 J27 J28
LP 92 LP 54 90 91 59 LP
J35 J36 J37 J38 J39 J40 J41 J42 J43 J44 J45 J46 J47 J48 J49 J50 J51 J52
LP LP 53 LP 53 57 50 32 49 74 LP 74 LP LP 37
A09 A10 A11 A12 A13 A14 A15 A16 A17 A18 A19 A20 A21 A22 A23 A24 A25 A26 A27 A28 A29 A30 A31 A32 A33 A34
A39 A40 A41 A42 A43 A44 A45 A46 A47 A48 A49 A50 A51 A52 A53 A54 A55 A56 A57 A58 A59 A60 A61
54 54 54 54 87 LP LP LP
A68 A69 A70 A71 A72 A73 A74 A75 A76



45,000 nt
90,000
135,000
180,000
315,000
360,000
405,000
450,000
16S
495,000
540,000
585,000
675,000
720,000
765,000
810,000
855,000
900,000
23s rRNA
16s rRNA
5s rRNA
tRNA
cp9
cp26
cellobiose
lp17
lp25
lp28-1
lp28-2
lp28-3
lp28-4
lp36
lp38
lp54


l kb
ding proteins
ries
ypothetical
Signal peptide
LP Lipoprotein
Fe+++ Transporter
GES region
94 Paralogous gene family
Authentic Frame Shift
23S
16S

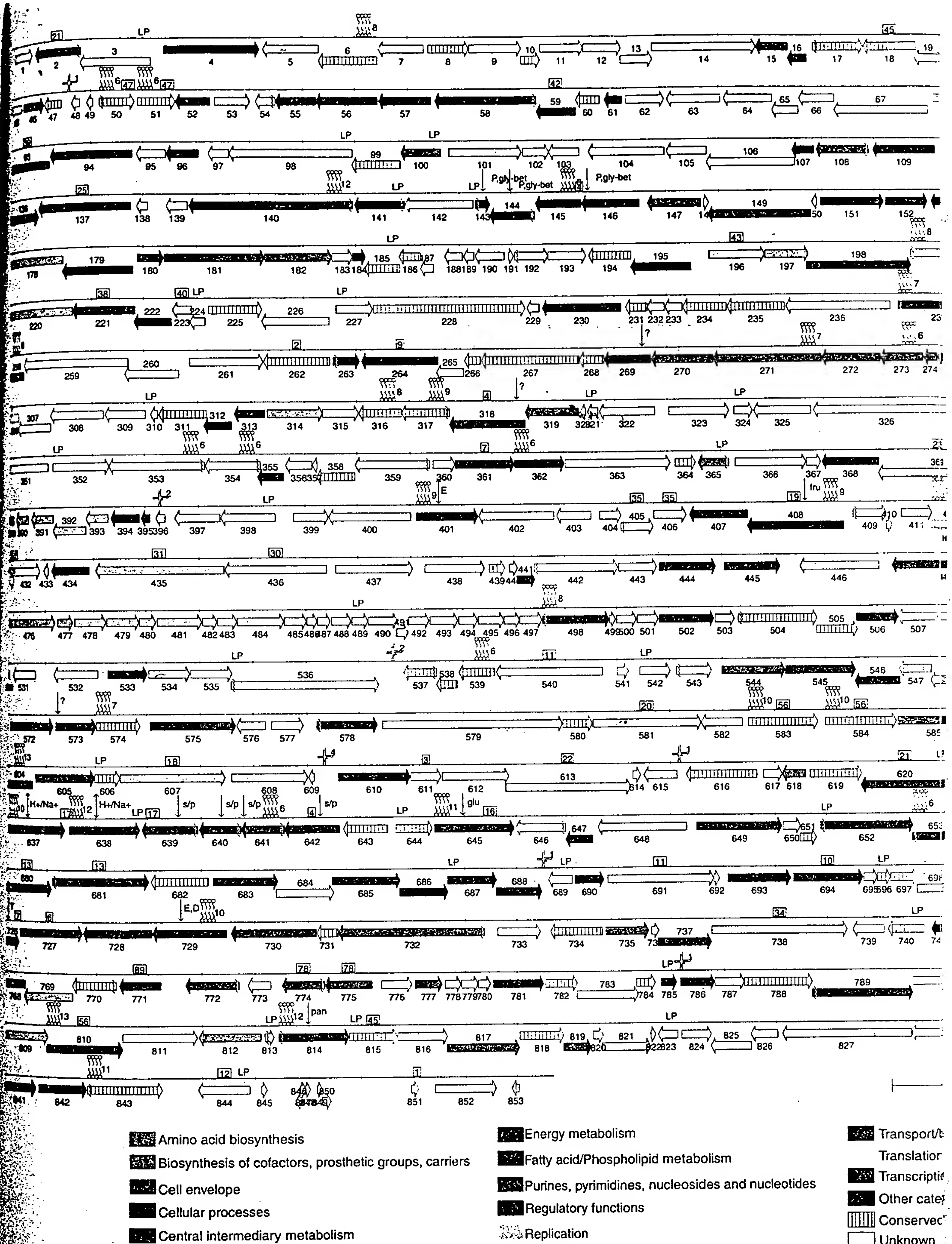

Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid/Phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
Translation
Conserved
Unknown

Table 2. Identification of *B. burgdorferi* genes

BB#	Identification (Species)	%Sim.
Amino acid biosynthesis		
Serine family		
BB601	serine OHMTase (glyA) (Ec)	73
Biosynthesis of cofactors, prosthetic groups, and carriers		
Folic acid		
BB026	methylenetetrahydrofolate DHase (folD) (Bs)	61
Heme and porphyrin		
BB197	protoporphyrinogen oxidase, put (Bs)	56
BB656	oxygen-independent coproporphyrinogen III oxidase, put (Bs)	51
Menaquinone and ubiquinone		
BB314	octaprenyl-diP Sase (ispB) (Ec)	57
Pantothenate		
BB812	pantothenate metabolism flavoprt (dfp) (Ec)	58
Pyridoxine		
BB768	pyridoxal kinase (pdxK) (Sc)	47
Thiamine		
BB621	4-methyl-5(b-OHethyl)-thiazole monoP biosyn prt (thiJ) (Ec)	58
Pyridine nucleotides		
BB522	NH(3)-dep NAD+ Sase (Rc)	65
Cell envelope		
Membranes, lipoproteins, and porins		
BB382	basic membrane prt B (bmpB) (Bb)	100
BB383	basic membrane prt A (bmpA) (Bb)	100
BB384	basic membrane prt C (bmpC) (Bb)	100
BB385	basic membrane prt D (bmpD) (Bb)	100
BB108	basic membrane prt (Tp)	50
BB319	exported prt (tpn38b) (Tp)	50
BB347	fibronectin/fibrinogen-BP, put (Sp)	53
BB442	inner membrane prt (Hi)	63
BB365	lipoprotein LA7 (Bb)	100
BB603	membrane-associated prt p66 (Bb)	100
BB753	membrane spanning prt, put (Se)	49
BB795	outer membrane prt (Ng)	48
BB167	outer membrane prt (tpn50) (Tp)	48
BB735	rare lipoprt A (rlpA) (Hi)	58
BB10	surface-located membrane prt 1 (lmp1) (Mh)	45
BB158	S2 prt (Bb)	60
lp54		
BBA24	decorin BP A (dbpA) (Bb)	94
BBA25	decorin BP B (dbpB) (Bb)	100
BBA36	lipoprotein (Bb)	54
BBA59	lipoprotein (Bb)	100
BBA62	lipoprotein (Bb)	100
BBA74	outer membrane porin (oms28) (Bb)	100
BBA15	outer surface prt A (ospA) (Bb)	99
BBA16	outer surface prt B (ospB) (Bb)	99
BBA03	outer membrane prt (Bb)	100
BBA52	outer membrane prt (Bb)	100
BBA05	S1 prt (Bb)	100
BBA04	S2 prt (Bb)	100
BBA60	surface lipoprt P27 (Bb)	81
lp38		
BBJ09	outer surface prt D (ospD) (Bb)	100
BBJ50	outer membrane prt, put (Bb)	61
BBJ51	vlsE1 prt, put (Bb)	62
BBJ52	vlsE1 prt, put (Bb)	55
lp36		
BBK53	outer membrane prt (Bb)	91
BBK52	prt p23 (Bb)	99
lp28-1		
BBF01	erpD prt, put (Bb)	53
BBF22	prt p23, put (Bb)	78
BBF32	vls recombination cassette Vls3-16 (Bb)	100
lp28-3		
BBH20	outer membrane porin (oms28), put (Bb)	74
BBH21	outer membrane porin (oms28), put (Bb)	74
lp25		
BBE09	prt p23 (Bb)	93
cp26		
BBB19	outer surface prt C (ospC)(Bb)	100
BBB07	outer surface prt, put (Bb)	45
cp9		
BBC06	exported prt A (eppA) (Bb)	100
Murein sacculus and peptidoglycan		
BB160	alanine racemase (alr) (Ec)	54
BB582	CPDase, put (Bs)	54
BB200	D-alanine-D-alanine ligase (ddlA) (Ec)	60
BB100	glutamate racemase (murI) (Ec)	56
BB625	N-Acmuramoyl-L-alanine amidase, put (Ef)	53
BB732	penicillin-BP (pbp-3) (Ng)	52
BB136	penicillin-BP (pbp-1) (Nm)	52
BB718	penicillin-BP (pbp-2) (Hi)	52
BB303	phospho-N-Acmuramoyl-pentapeptideTase (mraY) (Bb)	100
BB715	rod shape-determining prt (mreB-1) (Ec)	73
BB716	rod shape-determining prt (mreC) (Bs)	51
BB719	rod shape-determining prt (mreB-2) (Hi)	61
BB605	serine-type D-Ala-D-Ala CPDase (dacA) (Hi)	55
BB472	UDP-NAG 1-carboxy-vinylTase (murA) (Ec)	59
BB598	UDP-N-Acmuramate DHase (murB) (Bs)	55
BB817	UDP-N-Acmuramate-alanine ligase (murC) (Hi)	54
BB585	UDP-N-Acmuramoylalanine-D-glutamate ligase (murD) (Bs)	52
BB201	UDP-N-Acmuramoylalanyl-D-glutamate–2,6-diaminopimelate ligase (murE) (Hi)	53
BB304	UDP-N-Acmuramoylalanyl-D-glutamyl-2,6-diamino-pimelate-D-alanyl-D-alanine ligase (murF) (Bb)	100
BB767	UDP-N-Acglucosamine-N-Acmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol NAG Tase (murG) (Bs)	56
Surface polysaccharides, lipopolysaccharides and antigens		
BB744	antigen, p83/100 (Bb)	100
BB572	glycosyl Tase (lgtD) (Hi)	56
lp54		
BBA64	antigen, P35 (Bb)	100
BBA66	antigen, P35, put (Bb)	48
BBA73	antigen, P35, put (Bb)	52
lp38		
BBJ41	antigen, P35, put (Bb)	47
lp36		
BBK15	antigen, P35, put (Bb)	50
BBK50	immunogenic prt P37 (Bb)	99
BBK32	immunogenic prt P35, put (Bb)	97
BBK37	immunogenic prt P37, put (Bb)	71
BBK45	immunogenic prt P37, put (Bb)	49
BBK46	immunogenic prt P37, put (Bb)	68
BBK48	immunogenic prt P37, put (Bb)	49
lp28-3		
BBH32	antigen, P35, put (Bb)	55
lp28-4		
BBI36	antigen, P35, put (Bb)	48
lp25		
BBE31	antigen, P35, put (Bb)	54
Surface structures		
BB289	flgr assembly prt (fliH) (Bb)	100
BB291	flgr basal-body rod prt (fliF) (Bb)	100
BB293	flgr basal-body rod prt (flgC) (Bb)	100
BB294	flgr basal-body rod prt (flgB) (Bb)	100
BB774	flgr basal-body rod prt (flgG) (Sc)	68
BB271	flgr biosyn prt (flhA) (Bb)	100
BB272	flgr biosyn prt (flhB) (Bb)	100
BB273	flgr biosyn prt (fliR) (Bb)	100
BB274	flgr biosyn prt (fliQ) (Bb)	100
BB275	flgr biosyn prt (fliP) (Bb)	100
BB276	flgr biosyn prt (fliZ) (Bb)	100
BB147	flgr filament 41 kDa core prt (flaB) (Bb)	100
BB668	flgr filament outer layer prt (flaA) (Bb)	99
BB284	flgr hook assembly prt (flgD) (Bb)	100
BB283	flgr hook prt (flgE) (Bb)	100
BB181	flgr hook-associated prt (flgK) (Bb)	99
BB149	flgr hook-associated prt 2 (fliD) (Bb)	99
BB182	flgr hook-associated prt 3 (flgL) (Bb)	99
BB775	flgr hook-basal body complex prt (flhO) (Bs)	56
BB292	flgr hook-basal body complex prt (fliE) (Bb)	100
BB280	flgr motor rotation prt B (motB) (Bb)	100
BB281	flgr motor rotation prt A (motA) (Bb)	100
BB277	flgr motor switch prt (fliN) (Bb)	100
BB278	flgr motor switch prt (fliM) (Bb)	100
BB221	flgr motor switch prt (fliG-1) (Td)	54
BB290	flgr motor switch prt (fliG-2) (Bb)	100
BB772	flgr P-ring prt (flgI) (Ar)	51
BB279	flgr prt (fliL) (Bb)	100
BB282	flgr prt (flbD) (Bb)	99
BB285	flgr prt (flbC) (Bb)	100
BB286	flgr prt (flbB) (Bb)	100
BB287	flgr prt (flbA) (Bb)	100
BB180	flgr prt, put (Bb)	100
BB550	flgr prt (flaJ) (Vp)	57
BB270	flgr-associated GTP-BP (flhF) (Bb)	100
BB288	flgr-specific ATP Sase (fliI) (Bb)	100
Cellular processes		
General		
BB567	chemotaxis histidine Kase (cheA-1) (Bb)	99
BB669	chemotaxis histidine Kase (cheA-2) (Bb)	100
BB040	chemotaxis prt MTase (cheR-1) (Rs)	61
BB414	chemotaxis prt MTase (cheR-2) (Rm)	57
BB551	chemotaxis response regulator (cheY-1) (Tp)	74
BB570	chemotaxis response regulator (cheY-2) (Rm)	70
BB672	chemotaxis response regulator (cheY-3) (Bb)	98
BB660	GTP-BP (era) (Ec)	62
BB781	GTP-BP (obg) (Syn)	63
BB578	MC prt (mcp-1) (Tm)	52
BB596	MC prt (mcp-2) (Td)	57
BB597	MC prt (mcp-3) (Td)	52
BB680	MC prt (mcp-4) (Ec)	61
BB681	MC prt (mcp-5) (Ec)	52
BB415	prt-glutamate methylesterase (cheB-1) (Sc)	63
BB568	prt-glutamate methylesterase (cheB-2) (Sc)	59
BB312	purine-B chemotaxis prt (cheW-1) (Bb)	58
BB565	purine-B chemotaxis prt (cheW-2) (Ec)	66
BB670	purine-Bchemotaxis prt (cheW-3) (Bb)	100
Cell division		
BB058	cell division control prt 27, put (Mj)	46
BB195	cell division control prt 27, put (Mj)	54
BB361	cell division inhibitor, put (Bs)	53
BB299	cell division prt (ftsZ) (Bs)	72
BB789	cell division prt (ftsH) (Bs)	72
BB076	cell division prt, put (Ec)	71
BB257	cell division prt, put (Ec)	71
BB300	cell division prt (ftsA) (Bb)	100
BB302	cell division prt (ftsW) (Bb)	100
BB301	cell division prt (divIB) (Bb)	100
BB313	cell division prt (ftsJ) (Mj)	59
BB434	stage 0 sporulation prt J (spo0J) (Bs)	63

lp28-2

BBG08	stage 0 sporulation prt J (spoOJ) {Bb}	66

Cell killing

BB143	-hemolysin (hlyA) {Ah}	62
BB117	hemolysin III (yplQ) {Bs}	61
BB506	hemolysin (tlyA) {Sh}	59
BB059	hemolysin (tlyC) {Sh}	65
BB202	hemolysin, put {Syn}	54

Chaperones

BB741	chaperonin (groES) {Pg}	77
BB602	chaperonin, put {Cb}	72
BB519	grpE prt (grpE) {Bb}	100
BB295	heat shock prt (hslU) {Bb}	100
BB296	heat shock prt (hslV) {Bb}	100
BB649	heat shock prt (groEL) {Bb}	100
BB517	heat shock prt (dnaJ-1) {Bb}	100
BB655	heat shock prt (dnaJ-2) {Ca}	59
BB264	heat shock prt 70 (dnaK-1) {Bb}	61
BB518	heat shock prt 70 (dnaK-2) {Bb}	100
BB560	heat shock prt 90 (htpG) {Bb}	100

Detoxification

BB153	superoxide dismutase (sodA) {Hi}	68
BB690	neutrophil activating prt (napA) {Hi}	57
BB179	thiophene and furan oxidation prt (thdF) {Bb}	100

Protein and peptide secretion

BB154	preprt translocase sub (secA) {Bb}	100
BB395	preprt translocase sub (secE) {Bl}	62
BB498	preprt translocase sub (secY) {Sc}	64
BB362	prolipoprt diacylglyceryl Tase (lgt) {Ec}	56
BB652	prt-export membrane prt (secD) {Ec}	63
BB653	prt-export membrane prt (secF) {Hi}	63
BB030	signal peptidase I (lepB-1) {Bs}	51
BB031	signal peptidase I (lepB-2) {Syn}	57
BB263	signal peptidase I (lepB-3) {St}	57
BB469	signal peptidase II (lsp) {Sc}	60
BB694	signal recognition particle prt (ffh) {Bs}	70
BB610	trigger factor (tig) {Hi}	50

Transformation

BB591	competence locus E, put {Bs}	54
BB798	competence prt F, put {Hi}	52

Central intermediary metabolism

General

BB241	glycerol kinase (glpK) {Ec}	74
BB243	glycerol-3-P DHase, anaerobic (glpA) {Hi}	52
BB376	SAM Sase (metK) {Bs}	72

Amino sugars

BB152	glucosamine-6-P isomerase (nagB) {Hi}	79
BB151	N-Acglucosamine-6-P deAcase (nagA) {Hi}	54

Degradation of polysaccharides

BB620	-glucosidase, put {Syn}	58
BB002	-N-Achexosaminidase, put {As}	54

Phosphorus compounds

BB533	phnP prt (phnP) {Ec}	48

Polysaccharides - (cytoplasmic)

BB166	4- –glucanoTase (malQ) {Syn}	55
BB004	phosphoglucomutase (femD) {Mj}	52
BB835	phosphomannomutase (cpsG) {Hi}	57

Energy metabolism

Aerobic

BB728	NADH oxidase, water-forming (nox) {Sh}	59

Amino acids and amines

BB841	arginine deiminase (arcA) {Cp}	75
BB842	ornithine carbamoylTase (arcB) {Ng}	74

Anaerobic

BB016	glpE prt (glpE) {Hi}	53
BB087	L-lactate DHase (ldh) {Bs}	72

ATP-proton motive force interconversion

BB094	V-type ATPase, sub A (atpA) {Mb}	64
BB093	V-type ATPase, sub B (atpB) {Mb}	62
BB092	V-type ATPase, sub D (atpD) {Mj}	51
BB096	V-type ATPase, sub E (atpE) {Mj}	54
BB091	V-type ATPase, sub I (atpI) {Eh}	53
BB090	V-type ATPase, sub K (atpK) {Mj}	54

Electron transport

BB061	thioredoxin (trxA) {Ec}	59
BB515	thioredoxin RDase (trxB) {Bb}	99

Fermentation

BB622	acetate kinase (ackA) {Ec}	63
BB589	P AcTase (pta) {Tt}	65

Glycolysis

BB337	enolase (eno) {Bs}	79
BB445	fructose-bisP aldolase (fba) {Ec}	80
BB730	glucose-6-P isomerase (pgi) {Pf}	62
BB057	glyceraldehyde 3-P DHase (gap) {Bb}	99
BB630	1-phosphofructoKase (fruK) {Hi}	52
BB056	phosphoglycerate Kase (pgk) {Bb}	99
BB658	phosphoglycerate mutase (gpmA) {Ec}	79
BB348	pyruvate Kase (pyk) {Bs}	62
BB727	pyroP-fructose 6-P 1-PPTase (pfk) {Eh}	65
BB020	pyroP-fructose 6-P 1-PPTase, sub (pfpB) {Bb}	100
BB055	trioseP isomerase {Bb}	100

Pentose phosphate pathway

BB222	glucose-6-P 1-DHase, put {As}	48
BB636	glucose-6-P 1-DHase (zwf) {Hi}	64
BB561	phosphogluconate DHase (gnd) {Sd}	71
BB657	ribose 5-P isomerase (rpi) {Mj}	61

Sugars

BB407	mannose-6-P isomerase (manA) {Ec}	54
BB444	nucleotide sugar epimerase {Vc}	69
BB676	phosphoglycolate PPase (gph) {Hi}	50
BB207	UTP-glucose-1-P uridylylTase (gtaB) {Bs}	63
BB545	xylulokinase (xylB) {Bs}	43

Fatty acid and phospholipid metabolism

General

BB037	1-acyl-sn-glycerol-3-P AcTase (plsC) {Bb}	100
BB685	3-OH-3-methylglutaryl-CoA RDase (mvaA) {Pm}	52
BB683	3-OH-3-methylglutaryl-CoA Sase {At}	53
BB109	Ac-CoA C-AcTase (fadA) {Hi}	67
BB704	acyl carrier prt {Syn}	65
BB721	CDP-diacylglycerol-glycerol-3-P 3-phosphatidylTase {Bs}	55
BB327	glycerol-3-P O-acylTase, put {So}	50
BB368	glycerol-3-P DHase, NAD(P)+ (gpsA) {Bs}	54
BB137	long-chain-fatty-acid CoA ligase {Syn}	54
BB593	long-chain-fatty-acid CoA ligase {Syn}	56
BB688	melvalonate Kase {Mj}	51
BB686	mevalonate pyroP DCase {Sc}	52
BB119	phosphatidate cytidylylTase (cdsA), AFS{Ec}	61
BB249	phosphatidylTase {Hp}	52
BB687	phosphomevalonate Kase, put {Sc}	53

Purines, pyrimidines, nucleosides, nucleotides

Nucleotide and nucleoside interconversion

BB417	adenylate kinase (adk) {Bs}	64
BB128	cytidylate kinase (cmk-1) {Bs}	58
BB819	cytidylate kinase (cmk-2) {Mj}	57
BB463	nucleoside-diP kinase (ndk) {Bs}	70
BB793	thymidylate kinase (tmk) {Mj}	59
BB571	uridylate kinase (smbA) {Mj}	54

Purine ribonucleotide biosynthesis

BB544	phosphoribosyl pyroP Sase (prs) {Mp}	59

cp26

BBB18	GMP Sase (guaA) {Bb}	100
BBB17	IMP DHase (guaB) {Bb}	100

Pyrimidine ribonucleotide biosynthesis

BB575	CTP Sase (pyrG) {Mj}	71

Salvage of nucleosides and nucleotides

BB777	adenine phosphoribosylTase (apt) {Ta}	63
BB618	cytidine deaminase (cdd) {Mp}	61
BB239	deoxyguanosine/deoxyadenosine kinase(I) sub 2 (dck) {La}	59
BB375	pfs prt (pfs-1) {Ec}	64
BB588	pfs prt (pfs-2) {Hi}	59
BB791	thymidine kinase (tdk) {Bs}	47
BB015	uridine kinase (udk) {Bb}	100

lp36

BBK17	adenine deaminase (adeC) {Bs}	57

Regulatory functions

General

BB184	carbon storage regulator (csrA) {Hi}	63
BB647	ferric uptake regulation prt (fur) {Sp}	48
BB198	guanosine-3',5'-bis(diP) 3'-pyrophosphohydrolase (spoT) {Ec}	61
BB737	histidine phosphoKase/PPase, put {Ml}	49
BB176	methanol DHase regulator (moxR) {Bb}	99
BB416	pheromone shutdown prt (traB) {Ef}	61
BB042	P transport system regulatory prt (phoU) {Pa}	57
BB379	prt Kase C1 inhibitor (pkcI) {Bb}	100
BB419	response regulatory prt (rrp-1) {Syn}	57
BB763	response regulatory prt (rrp-2) {Ec}	67
BB764	sensory transduction histidine Kase, put {Bs}	60
BB420	sensory transduction histidine Kase, put {Syn}	61
BB693	xylose operon regulatory prt (xylR-1) {Th.}	48
BB831	xylose operon regulatory prt (xylR-2) {Syn}	51

lp54

BBA07	chpAI prt, put {Ec}	55

Replication

Degradation of DNA

BB411	endonuclease precursor (nucA) {As}	53

DNA replication, restriction, modification, recombination, and repair

BB422	3-methyladenine DNA glycosylase (mag) {At}	56
BB827	ATP-dep helicase (hrpA) {Ec}	61
BB437	chromosomal replication init prt (dnaA) {Bb}	100
BB435	DNA gyrase, sub A (gyrA) {Bs}	67
BB436	DNA gyrase, sub B (gyrB) {Bb}	99
BB344	DNA helicase (uvrD) {Ec}	55
BB552	DNA ligase (lig) {Ta}	56
BB211	DNA mismatch repair prt (mutL) {Hi}	55
BB797	DNA mismatch repair prt (mutS) {Hi}	57
BB098	DNA mismatch repair prt, put {Syn}	51
BB548	DNA polymerase I (polA) {Hi}	61
BB579	DNA polymerase III, sub (dnaE) {Ec}	62
BB438	DNA polymerase III, sub (dnaN) {Bb}	100
BB461	DNA polymerase III, sub / (dnaX) {Bs}	61
BB710	DNA primase (dnaG) {Bs}	56
BB581	DNA recombinase (recG) {Syn}	60
BB828	DNA topoisomerase I (topA) {Syn}	64
BB035	DNA topoisomerase IV (parC) {Bb}	58
BB036	DNA topoisomerase IV (parE) {Bb}	56
BB745	endonuclease III (nth) {Syn}	59
BB837	excinuclease ABC, sub A (uvrA) {Ec}	4
BB836	excinuclease ABC, sub B (uvrB) {Ec}	71
BB457	excinuclease ABC, sub C (uvrC) {Syn}	57
BB534	exodeoxyribonuclease III (exoA) {Bs}	67
BB632	exodeoxyribonuclease V, chain	

	(recD) {Ec}	54
BB633	exodeoxyribonuclease V, chain (recB) {Hi}	51
BB634	exodeoxyribonuclease V, chain (recC) {Hi}	51
BB829	exonuclease SbcD (sbcD) {Ec}	55
BB830	exonuclease SbcC (sbcC) {Ec}	52
BB177	glucose-inhibited div prt B (gidB) {Bb}	99
BB178	glucose-inhibited div prt A (gidA) {Bb}	100
BB022	Holliday junction DNA helicase (ruvB) {Bb}	100
BB023	Holliday junction DNA helicase (ruvA) {Bb}	100
BB014	primosomal prt N (priA) {Bb}	100
BB131	recA prt (recA) {Bb}	100
BB607	rep helicase, ss DNA-dep ATPase (rep) {Hi}	61
BB111	replicative DNA helicase (dnaB) {Ec}	58
BB114	ss DNA-BP (ssb) {Syn}	62
BB254	ss-DNA-specific exonuclease (recJ) {Hi}	52
BB623	transcription-repair coupling factor (mfd) {Hi}	60
BB053	uracil DNA glycosylase (ung) {Hi}	68

lp28-2

BBG32	replicative DNA helicase, put {Bs}	59

lp25

BBE29	adenine specific DNA MTase, put {Hp}	57

Transcription

General

BB052	spoU prt (spoU) {Ec}	54

Degradation of RNA

BB805	polyribonucleotide nucleotidylTase (pnpA) {Bs}	68
BB046	ribonuclease H (rnhB) {Hi}	66
BB705	ribonuclease III (rnc) {Bs}	62
BB441	ribonuclease P prt component (rnpA) {Bb}	100

DNA-dependent RNA polymerase

BB502	DNA-directed RNA polymerase (rpoA) {Bs}	64
BB389	DNA-directed RNA polymerase (rpoB) {Bb}	97
BB388	DNA-directed RNA polymerase (rpoC) {Ec}	71
BB771	RNA polymerase sigma factor (rpoS) {Pa}	61
BB712	RNA polymerase sigma-70 factor (rpoD) {Bb}	100
BB450	RNA polymerase sigma-54 factor (ntrA) {Av}	57

Transcription factors

BB107	N utilization substance prt B (nusB) {Ec}	62
BB800	N-utilization substance prt A (nusA) {Bs}	62
BB394	transcription antitermination factor (nusG) {Ec}	64
BB132	transcription elongation factor (greA) {Ec}	56
BB355	transcription factor, put {Mx}	47
BB230	transcription termination factor Rho (rho) {Bb}	100

RNA processing

BB706	polynucleotide adenylylTase (papS) {Bs}	57

Translation

General

BB590	dimethyladenosine Tase (ksgA) {Bs}	61
BB802	ribosome-B factor A (rbfA) {Bs}	62

Amino acyl tRNA synthetases

B220	alanyl-tRNA Sase (alaS) {Ec}	62
B594	arginyl-tRNA Sase (argS) {Mj}	55
B101	asparaginyl-tRNA Sase (asnS) {Ec}	73
B446	aspartyl-tRNA Sase (aspS) {Ec}	66
B599	cysteinyl-tRNA Sase (cysS) {Hi}	58
B372	glutamyl-tRNA Sase (gltX) {Rm}	63
B371	glycyl-tRNA Sase (glyS) {Ta}	68
B35	histidyl-tRNA Sase (hisS) {Mj}	59
BB833	isoleucyl-tRNA Sase (ileS) {Sc}	66
BB251	leucyl-tRNA Sase (leuS) {Bs}	70
BB659	lysyl-tRNA Sase {Mj}	54
BB587	methionyl-tRNA Sase (metG) {Sc}	67
BB514	phenylalanyl-tRNA Sase, sub (pheT) {Bb}	100
BB513	phenylalanyl-tRNA Sase, sub (pheS) {Bb}	100
BB402	prolyl-tRNA Sase (proS) {Sc}	65
BB226	seryl-tRNA Sase (serS) {Bs}	62
BB720	threonyl-tRNA Sase (thrZ) {Bs}	67
BB005	tryptophanyl-tRNA Sase (trsA) {Cl}	65
BB370	tyrosyl-tRNA Sase (tyrS) {Bs}	62
BB738	valyl-tRNA Sase (valS) {Bs}	67

Degradation of proteins, peptides, and glycopeptides

BB608	aminoacyl-histidine dipeptidase (pepD) {Hi}	55
BB366	aminopeptidase I (yscI) {Bb}	100
BB069	aminopeptidase II {Bs}	57
BB611	ATP-dep Clp protease proteolytic component (clpP-1) {Hi}	79
BB757	ATP-dep Clp protease proteolytic component (clpP-2) {Hi}	67
BB369	ATP-dep Clp protease, sub A (clpA) {Ec}	56
BB612	ATP-dep Clp protease, sub X (clpX) {Ec}	75
BB834	ATP-dep Clp protease, sub C (clpC) {Pp}	67
BB253	ATP-dep protease LA (lon-1) {Bb}	100
BB613	ATP-dep protease LA (lon-2) {Hi}	65
BB359	carboxyl-terminal protease (ctp) {Syn}	65
BB203	Lambda CII stability-governing prt (hflK) {Ec}	56
BB204	Lambda CII stability-governing prt (hflC) {Ec}	56
BB248	oligoendopeptidase F (pepF) {Ll}	58
BB067	peptidase, put {Sc}	56
BB104	periplasmic serine protease DO (htrA) {Hi}	60
BB430	proline dipeptidase (pepQ) {Hi}	49
BB769	sialoglycoprotease (gcp) {Hi}	60
BB627	vacuolar X-prolyl dipeptidyl aminopeptidase I (pepX) {Ml	55
BB118	zinc protease, put {Hi}	54
BB536	zinc protease, put {Hi}	52

Nucleoproteins

BB232	hbbU prt {Bb}	100

Protein modification

BB105	methionine aminopeptidase (map) {Bs}	68
BB065	polypeptide deformylase (def) {Syn}	67
BB648	serine/threonine kinase, put {Pf}	51

Ribosomal proteins: synthesis and modification

BB392	ribosomal prt L1 (rplA) {Bs}	71
BB481	ribosomal prt L2 (rplB) {Bb}	99
BB478	ribosomal prt L3 (rplC) {Bb}	99
BB479	ribosomal prt L4 (rplD) {Bb}	100
BB490	ribosomal prt L5 (rplE) {Hi}	80
BB493	ribosomal prt L6 (rplF) {Sc}	72
BB390	ribosomal prt L7/L12 (rplL) {Sc}	75
BB112	ribosomal prt L9 (rplI) {Ec}	57
BB391	ribosomal prt L10 (rplJ) {Bs}	61
BB393	ribosomal prt L11 (rplK) {Tm}	73
BB339	ribosomal prt L13 (rplM) {Hi}	72
BB488	ribosomal prt L14 (rplN) {Tm}	79
BB497	ribosomal prt L15 (rplO) {Bs}	68
BB485	ribosomal prt L16 (rplP) {Syn}	81
BB503	ribosomal prt L17 (rplQ) {Ec}	63
BB494	ribosomal prt L18 (rplR) {Bs}	69
BB699	ribosomal prt L19 (rplS) {Ec}	74
BB188	ribosomal prt L20 (rplT) {Ec}	70
BB778	ribosomal prt L21 (rplU) {Ec}	58
BB483	ribosomal prt L22 (rplV) {Bb}	100
BB480	ribosomal prt L23 (rplW) {Bb}	100
BB489	ribosomal prt L24 (rplX) {Ec}	64
BB780	ribosomal prt L27 (rpmA) {Hi}	82
BB350	ribosomal prt L28 (rpmB) {Ec}	62
BB486	ribosomal prt L29 (rpmC) {Bs}	65
BB496	ribosomal prt L30 (rpmD) {Bs}	60
BB229	ribosomal prt L31 (rpmE) {Bs}	69
BB703	ribosomal prt L32 (rpmF) {Bs}	62
BB396	ribosomal prt L33 (rpmG) {Bs}	76
BB440	ribosomal prt L34 (rpmH) {Bb}	100
BB189	ribosomal prt L35 (rpmI) {Ba}	74
BB499	ribosomal prt L36 (rpmJ) {Bs}	89
BB127	ribosomal prt S1 (rpsA) {Ec}	55
BB123	ribosomal prt S2 (rpsB) {Pa}	79
BB484	ribosomal prt S3 (rpsC) {Hi}	71
BB615	ribosomal prt S4 (rpsD) {Hi}	63
BB495	ribosomal prt S5 (rpsE) {Bs}	77
BB115	ribosomal prt S6 (rpsF) {Os}	50
BB386	ribosomal prt S7 (rpsG) {Sc}	75
BB492	ribosomal prt S8 (rpsH) {Syn}	66
BB338	ribosomal prt S9 (rpsI) {Hi}	71
BB477	ribosomal prt S10 (rpsJ) {Bb}	100
BB501	ribosomal prt S11 (rpsK) {Hi}	77
BB387	ribosomal prt S12 (rpsL) {An}	89
BB500	ribosomal prt S13 (rpsM) {Cp}	76
BB491	ribosomal prt S14 (rpsN) {Bs}	72
BB804	ribosomal prt S15 (rpsO) {Tt}	77
BB695	ribosomal prt S16 (rpsP) {Bs}	70
BB487	ribosomal prt S17 (rpsQ) {Mc}	76
BB113	ribosomal prt S18 (rpsR) {Bs}	78
BB482	ribosomal prt S19 (rpsS) {Bb}	99
BB233	ribosomal prt S20 (rpsT) {Bb}	100
BB256	ribosomal prt S21 (rpsU) {Mx}	68
BB516	rRNA methylase (yacO) {Mc}	66

tRNA modification

BB821	2-methylthio-N6-isopentyladenosine tRNA modification enzyme (miaA) {Ec}	53
BB084	AT (nifS) {Syn}	61
BB343	glu-tRNA amidoTase, sub C (gatC) {Bs}	56
BB341	glu-tRNA amidoTase, sub B (gatB) {Bs}	63
BB342	glu-tRNA amidoTase, sub A (gatA) {Bs}	61
BB064	methionyl-tRNA formylTase (fmt) {Ec}	56
BB787	peptidyl-tRNA hydrolase (pth) {Bb}	100
BB012	pseudouridylate Sase I (hisT) {Bb}	100
BB021	SAM: tRNA ribosylTase-isomerase {Bb}	96
BB809	tRNA-guanine transglycosylase (tgt) {Zm}	60
BB698	tRNA (guanine-N1)-MTase (trmD) {Mg}	68
BB803	tRNA pseudouridine 55 Sase (truB) {Ec}	57

Translation factors

BB088	GTP-B membrane prt (lepA) {Hi}	76
BB196	peptide chain release factor 1 (prfA) Hi}	73
BB074	peptide chain release factor 2 (prfB) {Sc}	70
BB121	ribosome releasing factor (frr) {Mt}	68
BB169	translation initiation factor 1 (infA) {Ec}	87
BB801	translation initiation factor 2 (infB) {Bs}	73
BB190	translation initiation factor 3 (infC) {Pv}	72
BB691	translation elongation factor G (fus-2) {Tm}	67
BB214	translation elongation factor P (efp) {Ec}	56
BB476	translation elongation factor TU (tuf) {Bb}	100
BB122	translation elongation factor TS (tsf) {Hi}	57
BB540	translation elongation factor G (fus-1) {Tm}	68

Transport and binding proteins

General

BB573	ABC transporter, ATP-BP {Bs}	53
BB742	ABC transporter, ATP-BP {Syn}	57
BB466	ABC transporter, ATP-BP {Hi}	74
BB754	ABC transporter, ATP-BP {Bl}	60
BB080	ABC transporter, ATP-BP {Mj}	63
BB269	ATP-BP (ylxH-1) {Bb}	100
BB726	ATP-BP (ylxH-2) {Bb}	54

lp38

BBJ26	ABC transporter, ATP-BP {Mj}	62

Amino acids, peptide, and amines

BB729	glutamate transporter (gltP) {Bs}	55
BB401	glutamate transporter, put {Bs}	53
BB146	GBP ABC transporter, ATP-BP	

	(proV) {Sc}	71
BB145	GBP ABC transporter, permease prt (proW) {Ec}	66
BB144	GBP ABC transporter, BP (proX) {Ec}	43
BB334	OP ABC transporter, ATP-BP (oppD) {Bs}	75
BB335	OP ABC transporter, ATP-BP (oppF) {Bs}	80
BB332	OP ABC transporter, permease prt (oppB-1){Ec}	68
BB747	OP ABC transporter, permease prt (oppB-2){Bs}	54
BB333	OP ABC transporter, permease prt (oppC-1){Hi}	64
BB746	OP ABC transporter, permease prt (oppC-2){Bs}	52
BB328	OP ABC transporter, periplasmic BP (oppA-1) {Bb}	74
BB329	OP ABC transporter, periplasmic BP (oppA-2) {Bb}	94
BB330	OP ABC transporter, periplasmic BP (oppA-3) {Bb}	81
BB642	SP ABC transporter, ATP-BP (potA) {Ec}	69
BB641	SP ABC transporter, permease prt (potB) {Ec}	65
BB640	SP ABC transporter, permease prt (potC) {Ec}	63
BB639	SP ABC transporter, periplasmic BP (potD) {Ec}	53

lp54

BBA34	OP ABC transporter, periplasmic BP (oppA-4) {Bc}	66

cp26

BBB16	OP ABC transporter, periplasmic BP (oppA) {Bb}	78

Anions

BB218	P ABC transporter, ATP-BP (pstB) {Pa}	74
BB216	P ABC transporter, permease prt (pstC) {Ec}	58
BB217	P ABC transporter, permease prt (pstA) {Syn}	63
BB215	P ABC transporter, periplasmic P-BP (pstS) {Syn}	48

Carbohydrates, organic alcohols, and acids

BB240	glycerol uptake facilitator (glpF) {Bs}	57
BB604	L-lactate permease (lctP) {Ec}	57
BB318	methylgalactoside ABC transporter, ATP-BP (mglA) {Hi}	54
BB814	pantothenate permease (panF) {Ec}	63
BB448	phosphocarrier prt HPr (ptsH-1) (Mg)	56
BB557	phosphocarrier prt HPr (ptsH-2) {Hi}	69
BB558	phosphoenolpyruvate-prt PPase (ptsI) {Sc}	65
BB408	PTS system, fru-specific IIABC (fruA-1) {Ec}	65
BB629	PTS system, fru-specific IIABC (fruA-2) {Ec}	68
BB559	PTS system, glu-specific IIA (crr) {Bb}	100
BB645	PTS system, glu-specific IIBC (ptsG) {Sc}	67
BB116	PTS system, mal/glu-specific IIABC (malX) {Ec}	56
BB677	RG ABC transporter, ATP-BP (mglA) {Mg}	68
BB678	RG ABC transporter, permease prt (rbsC-1) {Mg}	51
BB679	RG ABC transporter, permease prt (rbsC-2) {Mp}	52

cp26

BBB04	PTS system, cello-specific IIC (celB) {Bs}	62
BBB05	PTS system, cello-specific IIA (celC) {Bs}	61
BBB06	PTS system, cello-specific IIB (celA) {Bs}	73
BBB29	PTS system, glu-specific IIBC, put {Ec}	70

Cations

BB724	K+ transport prt (ntpJ) {Eh}	60
BB380	Mg2+ transport prt (mgtE) {Bb}	100
BB164	Na+/Ca+ exchange prt, put {Mj}	59
BB447	Na+/H+ antiporter (napA) {Eh}	57
BB637	Na+/H+ antiporter (nhaC-1) {Bf}	48
BB638	Na+/H+ antiporter (nhaC-2) {Hi}	50

Other

BB451	chromate transport prt, put {Mj}	58

Other categories

Adaptations and atypical conditions

BB237	acid-inducible prt (act206) {Rm}	45
BB786	general stress prt (ctc) {Bs}	51
BB785	stage V sporulation prt G {Bm}	74
BB810	virulence factor mviN prt (mviN) {Hi}	51

Colicin-related functions

BB766	colicin V production prt, put {Hi}	52
BB546	outer membrane integrity prt (tolA) {Hi}	44

Drug and analog sensitivity

BB140	acriflavine resistance prt (acrB) {Hi}	53
BB258	bacitracin resistance prt (bacA) {Ec}	56
BB586	femA prt (femA) {Se}	47
BB141	membrane fusion prt (mtrC) {Hi}	47

lp28-4

BBI26	multidrug-efflux transporter {Hp}	55

lp25

BBE22	pyrazinamidase/nicotinamidase (pncA) {Mt}	56

Transposon-related functions

lp38

BBJ05	transposase-like prt, put {Bb}	89

lp36

BBK25	transposase-like prt, put {Bb}	80

lp28-1

BBF18	transposase-like prt, put {Bb}	96
BBF19	transposase-like prt, put {Bb}	96

lp28-2

BBG05	transposase-like prt {Bb}	99

lp28-3

BBH40	transposase-like prt, put {Bb}	57

lp17

BBD20	transposase-like prt, put {Bb}	99
BBD23	transposase-like prt, put {Bb}	88

Unknown

BB528	aldose RDase, put {Bs}	57
BB684	carotenoid biosyn prt, put {Ss}	58
BB671	chemotaxis operon prt (cheX) {Bb}	99
BB250	dedA prt (dedA) {Ec}	54
BB168	dnaK suppressor, put {Ec}	53
BB508	GTP-BP {Tp}	59
BB219	gufA prt {Mx}	54
BB421	hydrolase {Hi}	58
BB524	inositol monoPPase {Hs}	47
BB454	lipopolysaccharide biosyn-related prt {Mj}	49
BB702	lipopolysaccharide biosyn-related prt {Hi}	62
BB045	P115 prt {Mh}	53
BB336	P26 {Bb}	100
BB363	periplasmic prt {Bb}	100
BB033	small prt (smpB) {Rp}	70
BB297	smg prt {Bb}	100
BB443	spoIIIJ-associated prt (jag) {Bs}	56

lp54

BBA76	thy1 prt (thy1) {Dd}	68

lp28-4

BBI06	pfs prt (pfs) {Ec}	59

cp9

BBC09	rev prt (rev) {Bb}	62
BBC10	rev prt (rev) {Bb}	66

coding sequence. Biological roles were assigned to 59% of the 853 ORFs using the classification scheme adapted from Riley[29] (Fig. 1), 12% of ORFs matched hypothetical coding sequences of unknown function from other organisms, and 29% were new genes. The average relative molecular mass (M_r) of the chromosome-encoded proteins in *B. burgdorferi* is 37,529 ranging from 3,369 to 254,242, values similar to those observed in other bacteria including *Haemophilus influenzae*[20] and *Mycoplasma genitalium*[21]. The median isoelectric point (pI) for all predicted proteins is 9.7.

Analysis of codon usage in *B. burgdorferi* reveals that all 61 triplet codons are used. When both AU- and GC-containing codons specify a single amino acid, there is a marked bias (from 2-fold to more than 20-fold, depending on the amino acid) in the use of AU-rich codons. The most frequently used codons are AAA (Lys, 8.1%), AAU (Asn, 5.9%), AUU (Ile, 5.9%), UUU (Phe, 5.7%), GAA (Glu, 5.0%), GAU (Asp, 4.2%) and UUA (Leu, 4.2%). The most common amino acids are Ile (10.6%), Leu (10.3%), Lys (10.2%), Ser (7.8%) and Asn (7.2%). The high value for Lys is in agreement with the median calculated isoelectric point of 9.7.

Plasmid analysis

Analysis of the nucleotide sequence and Southern analyses on *B. burgdorferi* DNA indicate that, in addition to the large linear chromosome, isolate B31 contains linear plasmids of the following approximate sizes: 56 kilobase pairs (kbp) (lp56), 54 kbp (lp54), four plasmids of 28 kbp (lp28-1, lp28-2, lp28-3 and lp28-4), 38 kbp (lp38), 36 kbp (lp36), 25 kbp (lp25) and 17 kbp (lp17); and circular plasmids of the following sizes: 9 kbp (cp9), 26 kbp (cp26) and five or six homologous plasmids of 32 kbp (cp32). These include all of the plasmids previously identified in this strain, but comparisons with other B31 cultures suggest that this isolate may have lost one 21 kbp linear and one or two 32 kbp circular plasmids during growth in culture since its original isolation[11–14,19,30]. The sequences of all plasmids were assembled as part of this project. However, the assembled sequences of the cp32 and related lp56 plasmids could not be determined with a high degree of confidence because of DNA sequence similarity among them (≥99% in several regions of 3,000–5,000 bp per plasmid)[13,16] (Table 1). Improved assembly strategies are being tested to achieve closure on these plasmids (G. Sutton, unpublished). Plasmid lp17 is identical to that of lp16.9 from Barbour *et al.*[15].

The 11 plasmids we have described contain a total of 430 putative ORFs with an average size of 507 bp; plasmid G+C content ranges from 23.1% to 32.3%. Only 71% of plasmid DNA represents predicted coding sequences, a value significantly lower than that on the chromosome. This indicates that average intergenic distances are greater in the plasmids than in the chromosome, and that many potential ORFs contain authentic frameshifts or stops (see E29, for example), suggesting that they are decaying genes not encoding functional proteins. Of the 430 plasmid ORFs, only 70 (16%) could be identified and these include membrane proteins such as OspA-D, decorin-binding proteins, the VlsE lipoprotein recombination cassette, and the purine ribonucleotide biosynthetic enzymes GuaA and GuaB. We found that 100 ORFs (23%) match other hypothetical proteins from plasmids in this and related strains of *B. burgdorferi*[15,16,31]; 10 ORFs (2.3%) match hypothetical proteins from species other than *Borrelia*; and 250 ORFs (58%) have no database match.

We found that 47 paralogous gene families containing from 2 to 12 members account for 39% (169 ORFs) of the plasmid-encoded genes with no known biological role (Fig. 1). Paralogue families 32 and 50, typified by previously identified *B. burgdorferi* plasmid genes cp32 orfC and cp8.3 orf2, respectively, have some similarities to proteins involved in replication, segregation and control of copy number in other bacterial systems[16,31]. Previous studies have reported examples of plasmid gene duplication, but the extent of

Table 1 Genome features in *Borrelia burgdorferi*

Chromosome	910,725 bp (28.6% G+C)
Coding sequences (93%)	
RNAs (0.7%)	
Intergenic sequence (6.3%)	
853 coding sequences	
500 (59%) with identified database match	
104 (12%) match hypothetical proteins	
249 (29%) with no database match	
Plasmids	
cp9	9,386 bp (23.6% GC)
cp26	26,497 bp (26.3% GC)
lp17	16,828 bp (23.1% GC)
lp25	24,182 bp (23.3% GC)
lp28-1	26,926 bp (32.3% GC)
lp28-2	29,771 bp (31.5% GC)
lp28-3	28,605 bp (25.1% GC)
lp28-4	27,329 bp (24.4% GC)
lp36	36,834 bp (26.8% GC)
lp38	38,853 bp (26.1% GC)
lp54	53,590 bp (28.1% GC)
Coding sequences (71%)	
Intergenic sequence (29%)	
430 coding sequences	
70 (16%) with identified database match	
110 (26%) match hypothetical proteins	
250 (58%) with no database match	
Ribosomal RNA	Chromosome coordinates
16S	444581-446118
23S	438590-441508
5S	438446-438557
23S	435334-438267
5S	435201-435312
Stable RNA	
tmRNA	46973-47335
mpB	750816-751175
Transfer RNA	
34 species (8 clusters, 14 single genes)	

*The telomeric sequences of the nine linear plasmids assembled as part of this study were not determined; estimation of the number of missing terminal nucleotides by restriction analysis suggests that less than 1,200 bp is missing in all cases. Comparisons with previously determined sequences of lp 16.9 and one terminus of lp28-1 indicate that 25, 60 and 1,200 bp are missing, respectively.

able 2 Gene identification numbers are listed with the prefix BB as in Fig. 2. Each gene entified is listed in its functional role category (adapted from Riley[29]). The percentage of milarity and a two-letter abbreviation for genus and species for the best match are also own. An expanded version of this table with additional information is available on the orld-Wide Web at http://www.tigr.org/tdb/mdb/bbdb/bbdb.htm. Abbreviations of gene mes are: Ac, acetyl; BP, binding protein; biosyn, biosynthesis; cello, cellobiose; CPDase, rboxypeptidase; Dcase, decarboxylase; DHase, dehydrogenase; flgr, flagellar/flagellum; , fructose; GBP, glycine, betaine, L-proline; glu, glucose; Kase, kinase; mal, maltose; MC-ethyl-accepting chemotaxis; MTase, methyltransferase; NAG, *N*-acetylglucosamine; OH, droxy; OP, oligopeptide; P, phosphate; PPTase, phosphotransferase; PPase, phosphase; prt, protein; put, putative; RDase, reductase; RG, ribose/galactose; SAM, *S*-adenosylethionine; Sase, synthetase/synthase; SP, spermidine/putrescine; ss, single-stranded; b, subunit; Tase, transferase.
brevation of genus and species are: Ah, *Aeromonas hydrophila*; Ar, *Agrobacterium diobacter*; Al, *Alteromonas* sp.; Ab, *Anabaena* sp.; An, *Anacystis nidulans*; At, *arabidopsis aliana*; Av, *Azotobacter vinelandii*; Bf, *Bacillus firmus*; Bl, *Cacillus licheniformis*; Bm, *cillus megaterium*; Bs, *Bacillus stearothermophilus*; Bs, *Bacillus subtilis*; Bb, *Borrelia rgdorferi*; Bc, *Borrelia coriaceae*; Bh, *Borrelia hermsii*; Ba, *Buchnera aphidicola*; Ca, *ostridium acetobutylicum*; Cl, *Clostridium longisporum*; Cp, *Clostridium perfringens*; Cg, *rynegacterium glutamicum*; Cb, *Coxiella burnetii*; Cp, *Cyanophora paradoxa*; Dd, *ctyostelium discoideum*; Ec, *Escherichia coli*; Eh, *Entamoeba histolytica*; Ec, *terobacter cloacae*; El, *Enterococcus faecalis*; Eh, *Enterococcus hirae*; Ha, *emophilus aegyptius*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Hs, *Homo piens*; La, *Lactobacillus acidophilus*; Ll, *Lactococcus lactis*; Li, *Leptospira interrogans ovar lai*; Mj, *Methanococcus jannaschii*; Mb, *Methanosarcina barkeri*; Ml, *cobacterium leprae*; Mt, *Mycobacterium tuberculosis*; Mc, *Mycoplasma capricolum*; *Mycoplasma genitalium*; Mh, *Mycoplasma hominis*; Mh, *Mycoplasma hyorhinis*; Mm, *coplasma mycoides*; Mp, *Mycoplasma pneumoniae*; Mx, *Myxococcus xanthus*; Ng, *esseria gonorrhoeae*; Nm, *Neisseria meningitidis*; Os, *Odontella sinensis*; Pt, *ramecium tetraurelia*; Pa, *Pediococcus acidilactici*; Pf, *Plasmodium falciparum*; Pg, *phyromonas gingivalis*; Pv, *Proteus vulgaris*; Pa, *Pseudomonas aeruginosa*; Pm, *udomonas mevalonii*; Pp, *Pseudomonas putida*; Rm, *Rhizobium meliloti*; Rc, *odobacter capsulatus*; Rs, *Rhodobacter sphaeroides*; Rp, *Rickettsia prowazekii*; Sc, *charomyces cerevisiae*; Sc, *Salmonella choleraesius*; St, *Salmonella typhimurium*; Sh, *pulina hyodysenteriae*; Sd, *Shigella dysenteriae*; So, *Spinacia oleracea*; Sc, *aphylococcus carnosus*; Se, *Staphylococcus epidermidis*; Sp, *Streptococcus pyogenes*; Sc, *Streptomyces coelicolor*; Ss, *Sulfolobus solfataricus*; Syn, *Synechococcus* sp.; Sp, *nechocystis* PCC6803; Tt, *Thermoanaerobacterium thermosaccharolyticum*; Tb, Thermophilic bacterium RT8.B4.; Ttv, *Thermoproteus tenax* virus; Tm, *Thermotoga maritima*; Tat, *ous aquaticus thermophilus*; Ta, *Thermus aquaticus*; Td, *Treponema denticola*; Tp, *ponema pallidum*; Ta, *Triticum aestivum*; Tb, *Trypanosoma brucei* mitochondrion; Vc, *cholerae*; Vp, *Vibrio parahaemolyticus*; Zm, *Zymomonas mobilis*.

this redundancy has become even more apparent with the complete sequence of these 11 plasmids from isolate B31. Moreover, a preliminary search of 221 putative ORFs from the cp32s and lp56 indicates that at least 50% display ≥70% amino-acid similarity to ORFs from the other 11 plasmids presented here (data not shown). Although plasmid-encoded genes have been implicated in infectivity and virulence[17–19], the biological roles of most of these genes are not known. The significance of the large number of paralogous plasmid-encoded genes is not understood. These proteins may be expressed differentially in tick and mammalian hosts, or may undergo homologous recombination to generate antigenic variation in surface proteins. This hypothesis is supported by the identification of 63 plasmid-encoded putative membrane lipoproteins (Fig. 1).

Several copies of a putative recombinase/transposase similar to IS891-like transposases were identified in the *B. burgdorferi* plasmids. Linear plasmid 28-2 contains one full-length copy of this gene. Although no inverted repeats were found on either side of the transposase, there is a putative ribosome-binding site several nucleotides upstream of the apparent start codon, and a stem–loop structure ($-27\,\mathrm{kcal\,mol^{-1}}$) 195 bp downstream of the stop codon in an area with no ORFs. This transposase might represent a functional gene important for the frequent DNA rearrangements that presumably occur in *Borrelia* plasmids. There are other partial or nearly complete copies of the transposase gene that contain frame-destroying mutations elsewhere in the genome: two copies on lp17, one on lp36, one on lp38, one on lp28-3, two on lp28-1, and one near the right end of the large linear chromosome.

Origin of replication

The replication mechanism for the linear chromosome and plasmids in *B. burgdorferi* is not yet known. Replication possibly begins at the termini, as has been proposed for the poxvirus hairpin telomeres[32], or may begin from a single origin somewhere along the length of the linear replicon. Of the genes on the linear chromosome, 66% are transcribed away from the centre of the chromosome (Fig. 1), similar to the transcriptional bias observed for the genomes of *M. genitalium*[21] and *M. pneumoniae*[33]. It has been suggested that bacterial genes are optimally transcribed in the same direction as that in which replication forks pass over them, particularly for highly transcribed genes[34,35]

Given the transcriptional bias observed in *B. burgdorferi*, it seems likely that the origin of replication is near the centre of the chromosome. Because bacterial chromosomal replication origins are usually near *dnaA*[36], it is intriguing to note that this gene (BB437) lies almost exactly at the centre of the linear *B. burgdorferi* chromosome[10,27]. A centrally initiated, bi-directional replication fork would be equidistant from the two chromosome ends, and replication would traverse the rRNA genes in the same direction as transcription.

An analysis of GC skew, (G − C)/(G + C) calculated in 10-kilobase (kb) windows across the chromosome, shows a clear break at the putative origin of replication. The GC-skew values are uniformly negative from 0 to 450 kb (minus strand), and uniformly positive (plus strand) from 450 kb to the end of the chromosome (Fig. 2). Additional evidence for the location of the origin of replication comes from our discovery of an octamer, TTGTTTTT, whose skewed distribution in the plus versus the minus strand of the chromosome matches the GC skew (Fig. 2). The biological significance of this octamer has not yet been determined, although it may be analogous to the Chi site in *Escherichia coli* that is implicated in *recBCD* mediated recombination. No GC skew was observed in any of the plasmids, although the heptamer ATTTTTT displays a skewed distribution in the plus versus the minus strand of lp2[?] that changes at the approximate midpoint of the plasmid (not shown).

Transcription and translation

Genes encoding the three subunits (α, β, β') of the core RNA polymerase were identified in *B. burgdorferi* along with σ^{70} and two alternative σ factors, σ^{54} and *rpoS*. The role and specificity of each of these σ factors in transcription regulation in *B. burgdorferi* are not known. The *nusA*, *nusB* and *rho* genes, which are involved in transcription elongation and termination, were also identified.

A region of the genome with a significantly higher G + C content (43%), located between nucleotides 434,000 and 447,000, contains the rRNA operon. As previously reported, the rRNA operon in *B. burgdorferi* contains a 16S rRNA–Ala-tRNA–Ile-tRNA–23S rRNA–5S rRNA–23S rRNA–5S rRNA[37,38]. All of the genes are present in the same orientation, except for that encoding Ile tRNA. Four unrelated genes, encoding 3-methyladenine glycosylase, hydrolyase and two with no database match, are also present in the rRNA operon. Three of these genes are transcribed in the same direction as the rRNAs.

We identified in the chromosome 31 tRNAs with specificity for all 20 amino acids (Fig. 1). These are organized into 7 clusters plus 13 single genes. All tRNA synthetases are present except glutaminyl tRNA-synthetase. A single glutamyl tRNA synthetase probably aminoacylates both $\mathrm{tRNA^{Glu}}$ and $\mathrm{tRNA^{Gln}}$ with glutamate followed by transamidation by Glu-tRNA amidotransferase, a heterotrimeric enzyme present in *B. burgdorferi* and several Gram-positive bacteria and archaea[39]. The lysyl-tRNA synthetase (LysS) in *B. burgdorferi* is a class I type that has no resemblance to any known bacterial or eukaryotic LysS, but is most similar to LysS from the archaea[40].

Replication, repair and recombination

The complement of genes in *B. burgdorferi* involved in DNA replication is smaller than in *E. coli*, but similar to that in *M. genitalium*[21]. Three ORFs have been identified with high homology to four of the ten polypeptides in the *E. coli* DNA polymerase III: α, β and γ, and τ. In *E. coli*, the γ and τ proteins are produced by programmed ribosomal frameshifting. This observation suggests that DNA replication in *B. burgdorferi*, like that in *M. genitalium*, is accomplished with a restricted set of genes. *B. burgdorferi* has one

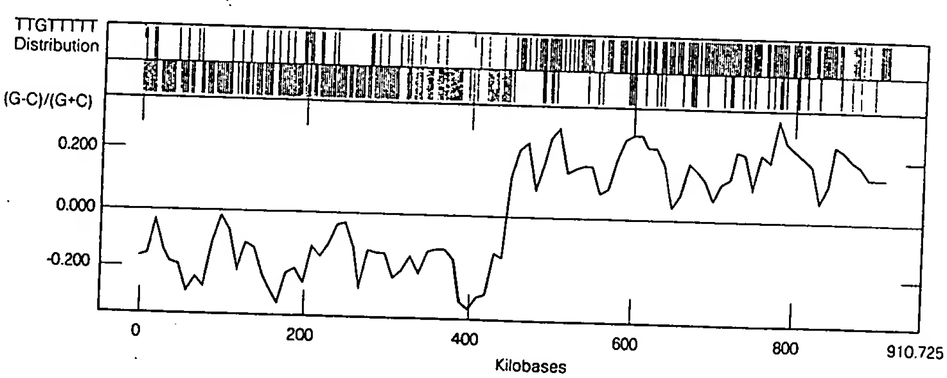
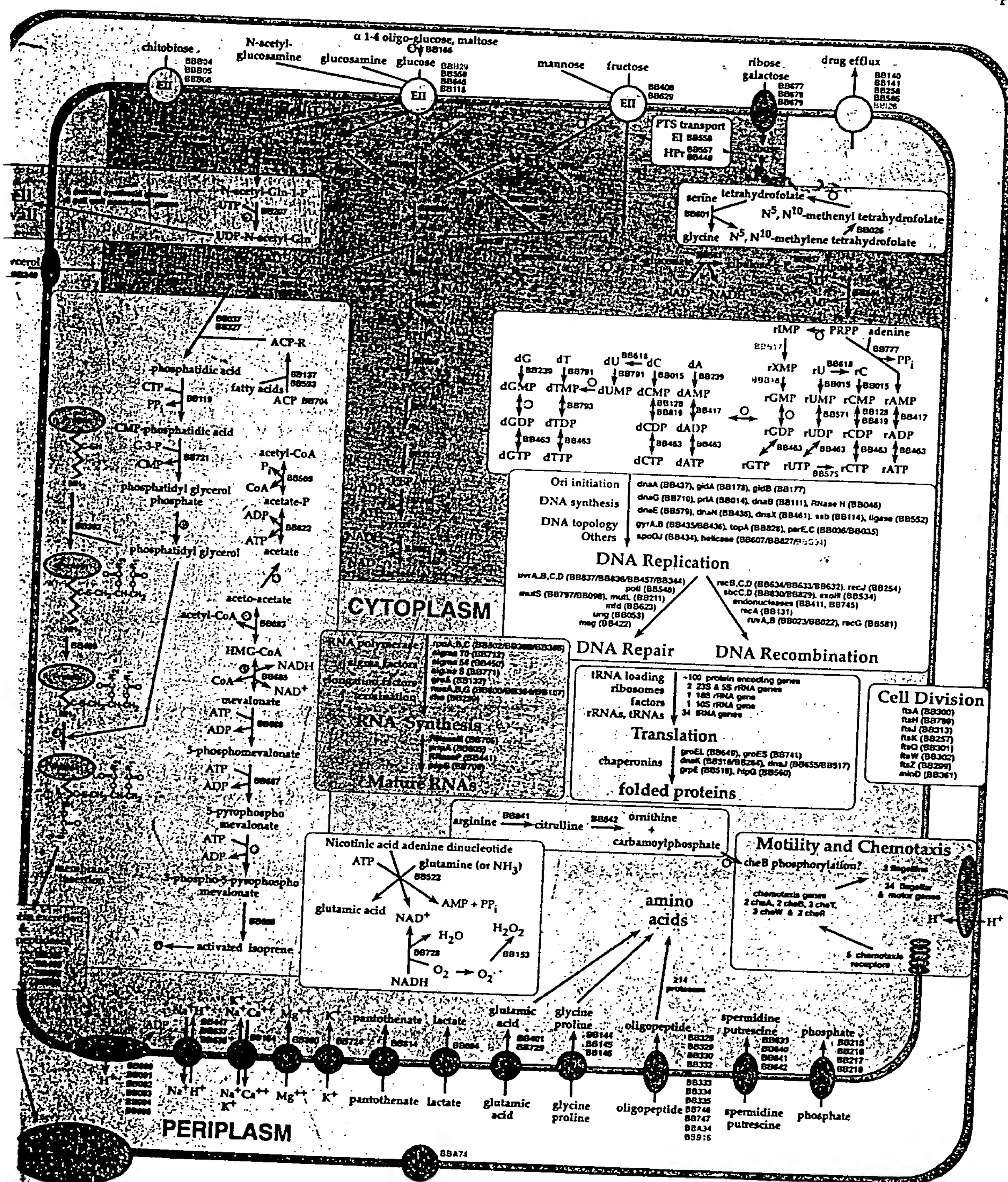


Figure 2 Distribution of TTGTTTTT and GC skew in the *B. burgdorferi* chromosome. Top, distribution of the octamer TTGTTTTT. The lines in the top panel represent the location of this octamer in the plus strand of the sequence, and those in the second panel represent the location of this oligomer in the minus strand of the sequence. Bottom, GC skew.

e I topoisomerase (*topA*) and two type II topoisomerases (gyrase l topoisomerase IV) for DNA topology management and chro-some segregation, despite its linear chromosomal structure. This gests that topoisomerase IV may be required for more than the uration of circular DNAs during segregation.

he DNA repair mechanisms in *B. burgdorferi* are similar to those *I. genitalium*. DNA excision repair can presumably occur by a :way involving endonuclease III, PolI and DNA ligase. The genes wo of three DNA mismatch repair enzyme (*mutS*, *mutL*) are present. The apparent absence of *mutH* is consistent with the lack of GATC (*dam*) methylation in strain B31 (S. Casjens, unpublished). Also present are genes for the repair of ultraviolet-induced DNA damage (*uvrA*, *uvrB*, *uvrC* and *uvrD*) (Table 2).

B. burgdorferi has a complete set of genes to perform homologous recombination, including *recA*, *recBCD*, *sbcC*, *sbcD*, *recG*, *ruvAB* and *recJ*. 3′-Exonuclease activity associated with *sbcB* in *E. coli* may be encoded by *exoA* (exodeoxynuclease III). Although *recA* is present, we found no evidence for *lexA*, which encodes the repressor that



transport and metabolic pathways in *B. burgdorferi*. A schematic *burgdorferi* cell providing an integrated view of the transporters mponents of the metabolism of this organism, as deduced from ed in the genome. The ORF numbers correspond to those listed in Table 2 (red indicates chromosomal and blue indicates plasmid ORFs). Presumed transporter specificity is indicated. Yellow circles indicate: places where particular uncertainties exist as to the substrate specificity, subcellular location or direction of catalysis: or expected activities that were not found.

regulates SOS genes in *E. coli.* No genes encoding DNA restriction or modification enzymes are present.

Biosynthetic pathways

The small genome size of *B. burgdorferi* is associated with an apparent absence of genes for the synthesis of amino acids, fatty acids, enzyme cofactors, and nucleotides, similar to that observed with *M. genitalium*[21] (Fig. 3, Table 2). The lack of biosynthetic pathways explains why growth of *B. burgdorferi in vitro* requires serum-supplemented mammalian tissue-culture medium. This is also consistent with previous biochemical data indicating that *Borrelia* lack the ability to elongate long-chain fatty acids, such that the fatty-acid composition of *Borrelia* cells reflects that present in the growth medium[6].

Transport

The linear chromosome of *B. burgdorferi* contains 46 ORFs and the plasmids contain 6 ORFs that encode transport and binding proteins (Fig. 3, Table 2). These gene products contribute to 16 distinct membrane transporters for amino acids, carbohydrates, anions and cations. The distribution of transporters between the four categories of functions in this section is similar to that observed in other heterotrophs (such as *Haemophilus influenzae, M. genitalium* and *H. pylori*), with most being dedicated to the import of organic compounds.

There are marked similarities between the transport capacity of *B. burgdorferi* and *M. genitalium.* Both genomes have a [...] number of recognizable transporters, so it is not clear how [...] can sustain diverse physiological reactions. Several of the [...] transporters in both genomes exhibit broad substrate spe[...] exemplified by the oligopeptide ABC transporter (*opp* oper[...] the glycine, betaine, L-proline transport system (*proVWX*) [...] fore, these organisms probably compensate for their [...] coding potential by producing proteins that can import a [...] variety of solutes. This is important because *B. burgdorferi* is [...] to synthesize any amino acids *de novo.* We were unable to iden[...] any transport systems for nucleosides, nucleotides, NAD/NADH [...] fatty acids, although they are likely to be present.

Glucose, fructose, maltose and disaccharides seem to be acqui[...] by the phosphoenolpyruvate:phosphotransferase system (PTS). The two nonspecific components, enzyme 1 (*ptsI*) and Hpr (*ptsH*) are associated in one operon with an apparently glucose-spe[...] phosphohistidine-sugar phosphotransferase enzyme IIA [...] Separate from this operon are four permeases (enzyme IIBC [...] *fruA* in two copies (fructose), *ptsG* (glucose) and *malX* (gluc[...] maltose) (Fig. 3, Table 2). The fructose-specific enzyme IIA [...] induced in the ORF with IIBC (*fruA*), as has been observed in *M. genitalium*[41]. Ribose may be imported by an ATP-binding cassette transporter (*rbsAC*). The *rbsAC* genes are transcribed in an opero[...] with a methyl-accepting chemotaxis protein that may respond to β-galactosides, suggesting that movement of the organisms toward sugars may be coupled to the transport process.

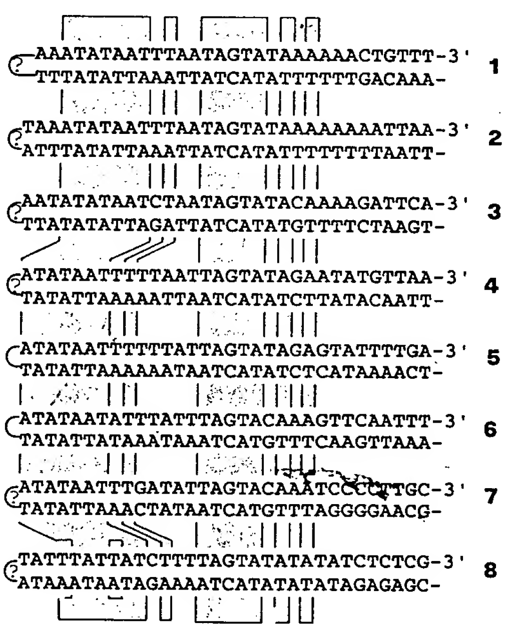


Figure 4 Telomere nucleotide sequences from *Borrelia* species. Nucleotide sequences are shown for known *Borrelia* telomeres as indicated: 1, *B. burgdorferi* Sh-2-82 chromosome left end; 2, *B. burgdorferi* B31 chromosome left end; 3, *B. afzelii* R-IP3 chromosome right end; 4, *B. burgdorferi* B31 chromosome right end; 5, *B. burgdorferi* B31 plasmid lp17 left end; 6, *B. burgdorferi* B31 plasmid lp17 right end; 7, *B. hermisii* plasmids bp7E and pb21E right ends; 8, *B. burgdorferi* B31 plasmid lp28-1 right end. In each case the telomere is at the left. Question marks (?) indicate locations where S1 nuclease was used to open terminal hairpins during the sequence determinations. Stippled areas highlight regions that appear to have been most highly conserved among these telomeres; no strong sequence conservation has been found near the right of the terminal 26 bp among the different sequences listed, except between the chromosomal left ends from strains B31 and Sh-2-82 (see text). The telomeric sequences of the strain B31 chromosome were determined in this report; the others are from references 14, 28, 30, 45, 46.

Energy metabolism

The limited metabolic capacity of *B. burgdorferi* is similar to that found in *M. genitalium* (Fig. 3, Table 2). Genes encoding all of the enzymes of the glycolytic pathway were identified. Analysis of the metabolic pathway suggests that *B. burgdorferi* uses glucose as a primary energy source, although other carbohydrates, including glycerol, glucosamine, fructose and maltose, may be used in glycolysis. Pyruvate produced by glycolysis is converted to lactate, consistent with the microaerophilic nature of *B. burgdorferi.* Generation of reducing power occurs through the oxidative branch of the pentose pathway. None of the genes encoding proteins of the tricarboxylic acid cycle or oxidative phosphorylation were identified. The similarity in metabolic strategies of two distantly related obligate parasites, *M. genitalium* and *B. burgdorferi,* suggests convergent evolutionary gene loss from more metabolically competent distant progenitors.

Addition of *N*-acetylglucosamine (NAG) to culture medium is required for growth of *B. burgdorferi*[6]. NAG is incorporated into the cell wall, and may also serve as an energy source. The cp26 plasmid encodes a PTS cellobiose transporter homologue that could have specificity for the structurally similar compound chitobiose (di-*N*-acetyl-D-glucosamine). A gene product on the chromosome with sequence similarity to chitobiase (BB2) may convert chitobiose to NAG. *B. burgdorferi* can metabolize NAG to fructose-6-phosphate, which then can enter the glycolytic cycle through the action of *N*-acetylglucosamine-6-phosphate deacetylase and glucosamine-6-phosphate isomerase. NAG is the primary constituent of chitin, which makes up the tick cuticle[6], and may be a source of carbohydrate for *B. burgdorferi* when it is associated with its tick host.

The parallels between *B. burgdorferi* and *M. genitalium* appear to extend to other aspects of their metabolism. Both organisms lack a respiratory electron transport chain, so ATP production must be accomplished by substrate-level phosphorylation. Consequently, membrane potential is established by the reverse reaction of the V_1V_0-type ATP synthase, here functioning as an ATPase to expel protons from the cytoplasm (Fig. 3, Table 2). The ATP synthase genes in *B. burgdorferi* appear to be transcribed as part of a seven-gene operon. They are not typical of those usually found in eubacteria, more closely resembling the eukaryotic vacuolar (V-type) and archaeal (A-type) H^+-translocating ATPases[42], both in si[...]

and sequence similarity, than the bacterial F_1F_0 ATPases. Genome analysis of *Treponema pallidum*, the pathogenic spirochaete that causes syphilis, has also revealed the presence of a V_1V_0-type ATP synthase (C. M. F. *et al.*, manuscript in preparation), suggesting that this may be a feature of spirochaetes.

Regulatory systems

Although the expression of *Borrelia* genes varies according to the current host species, temperature, host body location and other local factors, control of gene expression appears to differ from more well studied eubacteria. A typical set of homologues of heat-shock response genes is present (*groES, groEL, grpE, dnaJ, hslU, hslV, dnaK* and *htpG*), and *B. burgdorferi* is known to have such a response; however, it lacks the σ-32 that controls their transcription in *E. coli*. Only a few homologues to other eubacterial regulatory proteins are present, including only two response-regulator two-component systems.

Motility and chemotaxis

Like other spirochaetes, *B. burgdorferi* has periplasmic flagella that are inserted at each end of the cell and extend towards the middle of the cell body. The unique flagella allow the organism to move through viscous solutions, an ability that is presumed to be important in its migration to distant tissues following deposition in the skin layers[43]. Proteins involved in motility and chemotaxis are encoded by 54 genes, more than 6% of the *B. burgdorferi* chromosome, most of which are arranged in eight operons containing between 2 and 25 genes.

B. burgdorferi contains several copies of the chemotaxis genes (*cheR, cheW, cheA, cheY* and *cheB*) downstream of the methyl-accepting chemotaxis proteins. Other eubacteria also have duplications of some *che* genes, but those genes in *B. burgdorferi* are the most redundant set yet found. *B. burgdorferi* lacks recognizable virulence factors; thus, its ability to migrate to distant sites in the tick and mammalian host is probably dependent on a robust chemotaxis response. Multiple chemotaxis genes may provide redundancy in this system in order to meet such challenges or, alternatively, these genes may be differentially expressed under varied physiological conditions. Another speculative possibility is that the flagellar motors at the two ends of the *B. burgdorferi* cell are different and require different *che* systems. In support of this idea is the observation that one of the motor switch genes, *fliG*, is also present in two copies.

Membrane protein analysis

Much of the previous work on *B. burgdorferi* has focused on outer-surface membrane genes because of their potential importance in bacterial detection and vaccination. Nearly all *Borrelia* membrane proteins have been found to be typical bacterial lipoproteins. A search of *B. burgdorferi* ORFs for a consensus lipobox in the first 30 amino acids identified 105 putative lipoproteins, representing more than 8% of coding sequences. This contrasts with a total of only 20 putative lipoproteins in the 1.67-million base pair *H. pylori* genome (1% of coding sequences)[23]. The periplasmic binding proteins involved in transport of amino acids/peptides and phosphate in *B. burgdorferi* are candidate lipoproteins, suggesting that they may be anchored to the outer surface of the cytoplasmic membrane as in Gram-positive bacteria, rather than localized in the periplasmic space.

In better-characterized eubacteria, prolipoprotein diacylglycerol transferase (*lgt*), prolipoprotein signal peptidase (*lsp*), and apolipoprotein:phospholipid *N*-acyl transferase (*lnt*) are required for post-translational processing and addition of lipids to the amino-terminal cysteine. Genes for the first two of the enzymes (lgt and lsp) are present in the *B. burgdorferi* genome, but the gene for *lnt* was not found, although biochemical evidence argues for all three activities in *B. burgdorferi*[44]. The sequence similarity of an *lnt* homologue in *B. burgdorferi* may be too low to be identified using our search methods, or its activity may be present in a new enzyme. In *E. coli* the Sec protein export system moves lipoproteins through the inner membrane, and *Borrelia* carries a complete set of these protein-secretion gene homologues (*secA/D/E/F/Y* and *tth*; only the non-essential *secB* is missing).

Analysis of telomeres

The two chromosomal telomeres of strain B31 have similar 26-bp inverted terminal sequences (Fig. 4). We found no other similarity between the two ends, and these 26-bp sequences are very similar to the previously characterized *Borrelia* telomeres. Terminal restriction fragments from both B31 chromosomal termini were shown to exhibit snapback kinetics (data not shown), strongly indicating that both terminate in covalently closed hairpins, like previously characterized *Borrelia* telomeres[28,45,46].

The left chromosomal telomere of strain B31 is identical to the previously characterized left telomere of strain Sh-2-82 (ref. 28), except for a 31 bp insertion in B31 26 bp from the end. The right-most 7,454 bp contains surprisingly few ORFs, given the ORF density elsewhere on the chromosome. The function of this region is unknown, but it contains several unusual features. The right terminal 900 bp contains considerable homology to the left ends of lp17 and lp28-3. The region between 3,600 bp and 8,000 bp from the right end also contains several areas with similarity to plasmid sequences, including a portion of the transposase-like gene approximately 4,500 bp from the right end. The spacing between the two conserved motifs (ATATAAT and TAGTATA) in the right 26-bp terminal repeat is the same as most previously known plasmid telomeres but different from the previously known chromosomal telomeres. These findings support the idea that the right end of the *Borrelia* chromosome has historically exchanged telomeres with the linear plasmids[28].

Conclusions

The *B. burgdorferi* genome sequence will provide a new starting point for the study of the pathogenesis, prevention and treatment of Lyme disease. With the exception of a small number of putative virulence genes (haemolysins and drug-efflux proteins), this organism contains few, if any, recognizable genes involved in virulence or host–parasite interactions, suggesting that *B. burgdorferi* differs from better-studied eubacteria in this regard. It will be interesting to determine the role of the multi-copy plasmid-encoded genes, as previous work has implicated plasmid genes in infectivity and virulence. The completion of the genome sequence from a second spirochaete, *Treponema pallidum* (C.M.F. *et al.*, manuscript in preparation) will allow for the identification of genes specific to each species and to this bacterial phylum, and will provide further insight into prokaryotic diversity. □

Methods

Cell lines. A portion of a low-passage subculture of the original Lyme-disease spirochaete tick isolate[4] was obtained from A. Barbour. The type strain of *B. burgdorferi* (ATCC 35210)[3], B31, was derived from this isolate by limiting dilution cloning[5]. Cells were grown in Barbour–Stoenner–Kelly medium II (BSKII)[6], omitting the additions of antibiotics and gelatin, in tightly closed containers at 33–34 °C. Cells were subcultured three or fewer times *in vitro* between successive rounds of infection in C3H/HeJ mice to minimize loss of infectivity and plasmid content[17,18]. After four successive transfers of infection in mice, a primary culture of B31, established from infected ear tissue, was expanded to 2.5 l by four successive subcultures. All available evidence indicates that the B31 line used for preparation of genomic DNA was probably clonal, as genetic heterogeneity was undetectable by several criteria including macro-restriction analysis (S. Casjens, unpublished data) and plasmid analysis of clonal derivatives of the B31 line[13].

Sequencing. The *B. burgdorferi* genome was sequenced by a whole-genome random sequencing method previously applied to other microbial genomes[20–24].

An approximately 7.5-fold genome coverage was achieved by generating 19,078 sequences from a small insert plasmid library with an average edited length of 505 bases. The ends of 69 large insert lambda clones were sequenced to obtain a genome scaffold; 50% of the genome was covered by at least one lambda clone. Sequences were assembled using TIGR Assembler as described[20–24], resulting in a total of 524 assemblies containing at least two sequences, which were clustered into 85 groups based on linking information from forward and reverse sequence reads. All *Borrelia* sequences that had been mapped were searched against the assemblies in an attempt to delineate which were derived from the various elements of the *B. burgdorferi* genome. Some contigs were also located on the existing physical map by Southern analysis. Sequence and physical gaps for the chromosome were closed as described[20–24]. At the completion of the project, less than 3% of the chromosome had single-fold coverage. The linear chromosome of *B. burgdorferi* has covalently closed hairpin structures at its termini that are similar to those reported for linear plasmids in this organism[11]. The telomeric sequences (106 and 72 bp, respectively, from the left and right ends) were obtained after nicking the terminal loop with S1 nuclease and amplifying terminal sequences by ligation-mediated polymerase chain reaction (PCR) as described[28]. The unknown terminal sequence was determined in both directions on four independent plasmid clones of the amplified DNA from each telomere. A minimum amount of S1 nuclease was used and, because of their sequence similarity to other *Borrelia* telomeres, it is likely that few, if any, nucleotides were lost from the B31 chromosomal telomeres in this process.

Identification of ORFs. Coding regions (ORFs) were identified using compositional analysis using an interpolated Markov model based on variable-length oligomers[47]. ORFs of >600 bp were used to train the Markov model, as well as *B. burgdorferi* ORFs from GenBank. Once trained, the model was applied to the complete *B. burgdorferi* genome sequence and identified 953 candidate ORFs. ORFs that overlapped were visually inspected, and in some cases removed. Non-overlapping ORFs that were found between predicted coding regions and >30 amino acids in length were retained and included in the final annotation. All putative ORFs were searched against a non-redundant amino-acid database as described[20–24]. ORFs were also analysed using 527 hidden Markov models constructed for several conserved protein families (PFAM v2.0) using HMMER[48]. Families of paralogous genes were constructed by pairwise searches of proteins using FASTA. Matches that spanned at least 60% of the smaller of the protein pair were retained and visually inspected. A total of 94 paralogous gene families containing 293 genes were identified (Fig. 1).

Identification of membrane-spanning domains (MSDs). TopPred[49] was used to identify potential MSDs in proteins. A total of 526 proteins containing at least one putative MSD were identified, of which 183 were predicted to have more than one MSD. The presence of signal peptides and the probable position of a cleavage site in secreted proteins were detected using Signal-P as described[23]; 189 proteins were predicted to have a signal peptide. Lipoproteins were identified by scanning for a lipobox in the first 30 amino acids of every protein. A consensus sequence relaxed from that used for *H. pylori*[23] was defined for the purpose of this search based on known or putative *B. burgdorferi* lipoprotein consensus sequences.

Received 3 November; accepted 18 November 1997.

1. Steere, A. C. *et al.* Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three Connecticut communities. *Arthritis Rheum.* 20, 7 (1977).
2. Steere, A. C. Lyme disease. *New Engl. J. Med.* 321, 586–596 (1989).
3. Johnson, R. C., Schmidt, G. P., Hyde, F. W., Steigerwaalt, A. G. & Brenner, D. J. *Borrelia burgdorferi* sp. nov.: etiologic agent of Lyme disease. *Int. J. Syst. Bacteriol.* 34, 496–497 (1984).
4. Burgdorfer, W. *et al.* Lyme disease—a tick-borne spirochetosis? *Science* 216, 1317–1319 (1982).
5. Barbour, A. G., Burgdorfer, W., Hayes, S. F., Peter, O. & Aeschlimann, A. Isolation of a cultivable spirochete from *Ixodes ricinus* ticks of Switzerland. *Curr. Microbiol.* 8, 123–126 (1983).
6. Barbour, A. & Hayes, S. F. Biology of *Borrelia* species. *Microbiol. Rev.* 50, 381–400 (1986).
7. Baril, C., Richaud, C., Baranton, G. & Saint Girons, I. Linear chromosome of *Borrelia burgdorferi*. *Res. Microbiol.* 140, 507–516 (1989).
8. Ferdows, M. S. & Barbour, A. G. Megabase-sized linear DNA in the bacterium *Borrelia burgdorferi*, the Lyme disease agent. *Proc. Natl Acac Sci. USA* 86, 5969–5973 (1989).
9. Davidson, B., MacDougall, J. & Saint Girons, I. Physical map of the linear chromosome of the bacterium *Borrelia burgdorferi* 212, a causative agent of Lyme disease and localization of rRNA genes. *J. Bacteriol.* 174, 3766–3774 (1992).
10. Casjens, S. & Huang, W. M. Linear chromosome physical and genetic map of *Borrelia burgdorferi*, the Lyme disease agent. *Mol. Microbiol.* 8, 967–980 (1993).
11. Barbour, A. G. & Garon, C. F. Linear plasmids of the bacterium *Borrelia burgdorferi* have covalently closed ends. *Science* 237, 409–411 (1987).
12. Barbour, A. G. Plasmid analysis of *Borrelia burgdorferi*, the Lyme disease agent. *J. Clin. Microbiol.* 26, 475–478 (1988).
13. Casjens, S., van Vugt, R., Stevenson, B., Tilly, K. & Rosa, P. Homology throughout the [illegible] kilobase circular plasmids present in Lyme disease spirochetes. *J. Bacteriol.* 17, 217–[illegible]
14. Hinnebush, J. & Barbour, A. G. Linear- and circular-plasmid copy numbers in *Borrelia* [illegible] *J. Bacteriol.* 174, 5251–5257 (1992).
15. Barbour, A. G., Carter, C. J., Bundoc, V. & Hinnebush, J. The nucleotide sequence of a linear [illegible] of *Borrelia burgdorferi* reveals similarities to those of circular plasmids of other [illegible] *Bacteriol.* 178, 6625–6639 (1996).
16. Zuckert, W. R. & Meyer, J. Circular and linear plasmids of Lyme disease spirochetes share [illegible] homology: characterization of a repeated DNA element. *J. Bacteriol.* 178, 2287–2298 (1996).
17. Schwan, T. G., Burgdorfer, W. & Garon, C. Changes in infectivity and plasmid profile of [illegible] disease spirochete, *Borrelia burgdorferi*, as a result of *in vitro* cultivation. *Infect. Immun.* 56, [illegible] (1988).
18. Xu, Y., Kodner, C., Coleman, L. & Johnson, R. C. Correlation of plasmids with infectivity of [illegible] *burgdorferi sensu stricto* type strain B31. *Infect. Immun.* 64, 3870–3876 (1996).
19. Norris, S. J., Howell, J. K., Garza, S. A., Ferdows, M. S. & Barbour, A. G. High- and low-[illegible] phenotypes of clonal populations of *in vitro*-cultured *Borrelia burgdorferi*. *Infect. Immun.* 63, [illegible] 2212 (1995).
20. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512 (1995).
21. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–[illegible] (1995).
22. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1072 (1996).
23. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Heliocobacter pylori*. *Nature* 388, 539–547 (1997).
24. Klenk, H.-P. *et al.* The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390, 364–370 (1997).
25. Casjens, S., Ley, H., DeLange, M., Rosa, P. & Huang, W. Linear chromosomes of Lyme disease spirochetes; genetic diversity and conservation of gene order. *J. Bacteriol.* 177, 2769–2780 (1995).
26. Casjens, S. & Huang, W. in *Bacterial Genomes: Physical Structure and Analysis* (eds de Bruijn, [illegible] Lupski, J. & Weinstock, G. J.) (Chapman and Hall, New York, in the press)
27. Old, I., MacDougall, J., Saint Girons, I. & Davidson, B. Mapping of genes on the linear chromosome of the bacterium *Borrelia burgdorferi*: possible locations for its origin of replication. *FEMS Microbiol. Lett.* 99, 245–250 (1992).
28. Casjens, S. *et al.* Telomeres of the linear chromosomes of Lyme disease spirochetes: nucleotide sequence and possible exchange with linear plasmid telomers. *Mol. Microbiol.* 26, 581–596 (1997).
29. Riley, M. Functions of gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952 (1993).
30. Zhang, J., Hardham, J., Barbour, A. & Norris, S. Antigenic variation in Lyme disease *Borreliae* by promiscuous recombination of VMP-like sequence cassettes. *Cell* 89, 275–285 (1997).
31. Dunn, J. J. *et al.* Complete nucleotide sequence of a circular plasmid from the Lyme disease spirochete *Borrelia burgdorferi*. *J. Bacteriol.* 176, 2706–2717 (1994).
32. Traktman, P. in *DNA Replication in Eukaryotic Cells* (ed. DePamphilis, M.) 775–798 (Cold Spring Harbor Laboratory Press, NY, 1996).
33. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24, 4420–4449 (1996).
34. Brewer, B. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* 53, 679–686 (1988).
35. French, S. Consequences of replication fork movement through transcription units *in vivo*. *Science* 258, 1362–1365 (1992).
36. Ogasawara, N. & Yoshikawa, H. Genes and their organization in the replication origin region of the bacterial chromosome. *Mol. Microbiol.* 6, 629–634 (1992).
37. Gazumyan, A., Schwartz, J. J., Liveris, D. & Schwartz, I. Sequence analysis of the ribosomal RNA operon of the Lyme disease spirochete, *Borrelia burgdorferi*. *Gene* 146, 57–65 (1994).
38. Ojaimi, C., Davidson, B., Saint Girons, I. & Old, I. Conservation of gene arrangement and an unusual organization of rRNA genes in the linear chromosomes of Lyme disease *spirochaetes Borrelia burgdorferi*, *B. garinii* and *B. afzelli*. *Microbiology* 140, 2931–2940 (1994).
39. Cumow, A. W. *et al.* Glu-tRNA^{Gln} amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl Acad. Sci. USA* 94, 11819–11826 (1997).
40. Ibba, M., Bobo, J. L., Rosa, P. A. & Soll, D. Archaeal-type lysyt-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA* (in the press).
41. Reizer, J., Paulsen, I. T., Reizer, A., Titgemeyer, F. & Saier, M. H. Jr Novel phosphotransferase system genes revealed by bacterial genome analysis: The complete complement of *pts* genes in *Mycoplasma genitalium*. *Microb. Comp. Genom.* 1, 151–164 (1996).
42. Takase, K. *et al.* Sequencing and characterization of the *ntp* gene gluster for vacuolar-type Na$^+$-translocating ATPase of *Enterococcurs hirae*. *J. Biol. Chem.* 269, 11037–11044 (1994).
43. Sadziene, A., Rosa, P. A., Thompson, P. A., Hogan, D. M. & Barbour, A. G. Antibody-resistant mutants of *Borrelia burgdorferi*: *in vitro* selection and characterization. *J. Exp. Med.* 176, 799–809 (1992).
44. Brandt, M. E., Riley, B. S., Radolf, J. D. & Norgard, M. V. Immunogenic integral membrane proteins of *Borrelia burgdorferi* are lipoproteins. *Infect. Immun.* 58, 983–991 (1990).
45. Hinnebusch, J., Bergstrom, S. & Barbour, A. Cloning and sequence analysis of linear plasmid telomeres of the bacterium *Borrelia burgdorferi*. *Mol. Microbiol.* 4, 811–820 (1990).
46. Kitten, T. & Barbour, A. Juxtaposition of expressed variable antigen genes with a conserved telomere in the bacterium *Borrelia hermsii*. *Proc. Natl Acad. Sci. USA* 87, 6077–6081 (1990).
47. Salzberg, S., Delcher, A., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* (in the press).
48. Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein families based on seed alignments. *Proteins* 28, 405–420 (1997).
49. Claros, M. G. & von Heijne, G. TopPred II: an improved software for membrane proteins structure predictions. *Comput. Appl. Biosci.* 10, 685–686 (1994).

Acknowledgements. We thank A. G. Barbour for isolation of the *Borrelia burgdorferi* strain; A. Barbour, P. Rosa, K. Tilly, J. Riberio, B. Stevenson and D. Soll for discussions; N. K. Patel for technical assistance; M. Heaney, J. Scott and A. Saeed for software and database support; and V. Sapiro, B. Vincent and D. Maas for computer system support. This work was supported by a grant to J.C.V. and C.M.F. from the G. Harold and Leila Y. Mathers Charitable Foundation.

Correspondence and requests for materials should be sent to C.M.F. (e-mail: gbb@tigr.org). The annotated genome sequence and gene family alignments are available on the World-Wide Web at http://www.tigr.org/tdb/mdb/bbdb/bbdb.html. Sequences have been deposited with GenBank under the following accession numbers: AE00783 (chromosome); AE00784 (lp28-3); AE000785 (lp25); AE00786 (lp28-2); AE00787 (lp38); AE00788 (lp36); AE00789 (lp28-4); AE00790 (lp54); AE00791 (cp9); AE00792 (cp26); AE00793 (lp17); and AE00794 (lp28-1).

Appendix R

NCBI **BLAST Search Results** BLAST Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

You have searched a database generously provided by the Institute for Genomic Research (TIGR). Their Policy on Early Data Release is:

The Institute for Genomic Research (TIGR) releases data very rapidly to ensure that our scientific colleagues have access to information that may assist them in the search for genes and their biological function. Data releases do not constitute scientific publication, but rather provide investigators with information that may "jump-start" biological experimentation. Users of this information are encouraged to share their results with TIGR in order to improve annotation of the sequence data. Data or information may contain errors or be incomplete and should be regarded as preliminary.

TIGR asks that you acknowledge the source of information obtained from this site in any publication by including the following sentence in both the Materials and Methods and Acknowledgement sections: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" Also include the following text in the Acknowledgements, if applicable: "Sequencing of [organism name] was accomplished with support from [funding agency]." The name of the funding agency for each TIGR project can be found at http://www.tigr.org/tdb/mdb/mdb.html

Similarly, if you display this data or any information derived from it on a Web page, we ask that you prominently display the following notice on that webpage: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" We request that you notify us of your electronic presentation by sending email to www@tigr.org.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query=
(334 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

Sequences producing significant alignments:	Score (bits)	E Value

```
gb|AE000783|AE000783 Borrelia burgdorferi complete genome                         47   8e-07

 gb|AE000783|AE000783 Borrelia burgdorferi complete genome
             Length = 910724

 Score = 47.3 bits (110), Expect = 8e-07
 Identities = 32/149 (21%), Positives = 59/149 (39%)
 Frame = +2

Query: 12     EKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAG 71
              E L  S +   +  +A +      G+G  +    A +R L C+        CG C  C+ ++
Sbjct: 482678 ETLKHSIEKNKIANAYIFSGPRGVGKTSSARAFARCLNCRNGPTVMPCGECSNCKSIEND 482857

Query: 72     THPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXX 131
              +  D    +       G +   V  +R++ E++        +     ++.  +
Sbjct: 482858 SSLDVVEI---DGASNTSVQDIRQIKEEIMFPPAISKYRIYIIDEVHMLSNSAFNALLKT 483028

Query: 132    XEEPPAETWFFLATREPERLLATLRSRCR 160
               EEPP      F  AT E  +L  T++SRC+
Sbjct: 483029 IEEPPNYIVFIFATTESHKLPETIKSRCQ 483115


 Score = 23.9 bits (50), Expect = 8.7
 Identities = 13/32 (40%), Positives = 19/32 (58%)
 Frame = +3

Query: 267    NVDVPGLVAELANHLSPSRLQAILGDVCHIRE 298
              N D+    + +L N+LSPS L     GDV  ++E
Sbjct: 778332 NFDIFYELVKLRNNLSPSTLIIGNGDVLSLKE 778427


CPU time:     0.02 user secs.      0.06 sys. secs        0.08 total secs.

  Database: Borrelia burgdorferi
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 1,229,458
  Number of sequences in database:  12

Lambda     K      H
   0.322    0.137    0.00

Gapped
Lambda     K      H
   0.270    0.0470 4.94e-324


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 350020
Number of Sequences: 12
Number of extensions: 3681
Number of successful extensions: 16
Number of sequences better than 10.0: 2
Number of HSP's better than 10.0 without gapping: 2
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 10
Number of HSP's gapped (non-prelim): 5
length of query: 334
length of database: 409,819
effective HSP length: 38
effective length of query: 296
```

```
effective length of database: 409363
effective search space: 121171448
effective search space used: 121171448
frameshift window, decay const: 50,  0.1
T: 13
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 41 (21.9 bits)
S2: 50 (23.9 bits)
```

The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria

Siv G. E. Andersson*, Alireza Zomorodipour*, Jan O. Andersson*, Thomas Sicheritz-Pontén*, U. Cecilia M. Alsmark*, Raf M. Podowski*, A. Kristina Näslund*, Ann-Sofie Eriksson*, Herbert H. Winkler† & Charles G. Kurland*

* *Department of Molecular Biology, University of Uppsala, Uppsala S-75124, Sweden*
† *Department of Microbiology and Immunology, University of South Alabama, Mobile, Alabama 36688, USA*

We describe here the complete genome sequence (1,111,523 base pairs) of the obligate intracellular parasite *Rickettsia prowazekii*, the causative agent of epidemic typhus. This genome contains 834 protein-coding genes. The functional profiles of these genes show similarities to those of mitochondrial genes: no genes required for anaerobic glycolysis are found in either *R. prowazekii* or mitochondrial genomes, but a complete set of genes encoding components of the tricarboxylic acid cycle and the respiratory-chain complex is found in *R. prowazekii*. In effect, ATP production in *Rickettsia* is the same as that in mitochondria. Many genes involved in the biosynthesis and regulation of biosynthesis of amino acids and nucleosides in free-living bacteria are absent from *R. prowazekii* and mitochondria. Such genes seem to have been replaced by homologues in the nuclear (host) genome. The *R. prowazekii* genome contains the highest proportion of non-coding DNA (24%) detected so far in a microbial genome. Such non-coding sequences may be degraded remnants of 'neutralized' genes that await elimination from the genome. Phylogenetic analyses indicate that *R. prowazekii* is more closely related to mitochondria than is any other microbe studied so far.

The *Rickettsia* are α-proteobacteria that multiply in eukaryotic cells only. *R. prowazekii* is the agent of epidemic, louse-borne typhus in humans. Three features of this endocellular parasite deserve our attention. First, *R. prowazekii* is estimated to have infected 20–30 million humans in the wake of the First World War and killed another few million following the Second World War (ref. 1). Because it is the descendent of free-living organisms[2–4], its genome provides insight into adaptations to the obligate intracellular lifestyle, with probable practical value. Second, phylogenetic analyses based on sequences of ribosomal RNA and heat-shock proteins indicate that mitochondria may be derived from the α-proteobacteria[5,6]. Indeed, the closest extant relatives of the ancestor to mitochondria seem to be the *Rickettsia*[7–10]. That modern *Rickettsia* favour an intracellular lifestyle identifies these bacteria as the sort of organism that might have initiated the endosymbiotic scenario leading to modern mitochondria[11]. Finally, the genome of *R. prowazekii* is a small one, containing only 1,111,523 base pairs (bp). Its phylogenetic placement and many other characteristics identify it as a descendant of bacteria with substantially larger genomes[2–4]. Thus *Rickettsia*, like mitochondria, are good examples of highly derived genomes, the products of several types of reductive evolution.

The genome sequence of *R. prowazekii* indicates that these three features may be related. For example, prokaryotic genomes evolving within a cell dominated by a much larger, eukaryote genome and constrained by bottle-necked population dynamics will tend to lose genetic information[12,13]. Predictable sets of expendable genes will tend to disappear from the prokaryotic genome when they are made redundant by the activities of nuclear genes. Likewise, non-essential sequences and otherwise highly conserved gene clusters may be obliterated by deleterious mutations that are fixed in clonal parasite or organelle populations because they cannot be eliminated by selection. This process is ongoing in the *Rickettsia* genomes, as shown by the identification of sequences that have recently become pseudogenes. Also, a large fraction (~25%) of non-coding sequences in this genome may be gene remnants that have been degraded by mutation and have not yet been removed from the genome. Finally, transfer of genes from a mitochondrial ancestor to the nucleus of the host would both reduce the mitochondrial genome size and stabilize the symbiotic relationship. Phylogenetic reconstructions that identify genes in the *Rickettsia* genome as sister clades to eukaryotic homologues found in the nucleus or the organelle support this interpretation. *Rickettsia* and mitochondria probably share an α-proteobacterial ancestor and a similar evolutionary history.

General features of the genome

The circular chromosome of *R. prowazekii* strain Madrid E has 1,111,523 bp and an average G+C content of 29.1% (Figs 1, 2). The genome contains 834 complete open reading frames with an average length of 1,005 bp. Protein-coding genes represent 75.4% of the genome and 0.6% of the genome encodes stable RNA. We have assigned biological roles to 62.7% of the identified genes and pseudogenes; 12.5% of the identified genes match hypothetical coding sequences of unknown function and the remaining 24.8% represent unusual genes with no similarities to genes in other organisms (Table 1). Multivariate statistical analysis has shown that there is no major variation in codon-usage patterns among genes that are expressed in different amounts, indicating that codon-usage patterns in *R. prowazekii* may be dominated mainly by mutational forces[14]. G+C-content values at the three codon positions average 40.4, 31.2 and 18.6%, and these values are similar at different positions in the genome. We classified the open reading frames with significant sequence-similarity scores to gene sequences in the public databases into functional categories (Table 1) that allow comparisons with the metabolic profiles of other bacterial genomes[15–23].

Non-coding DNA. The coding content of previously sequenced bacterial genomes is, on average, 91%, ranging from 87% in *Haemophilus influenzae* to 94% in *Aquifex aeolicum*. In comparison, a large fraction of the *R. prowazekii* genome, 24%, represents non-coding DNA (Fig. 3). A small fraction of this corresponds to

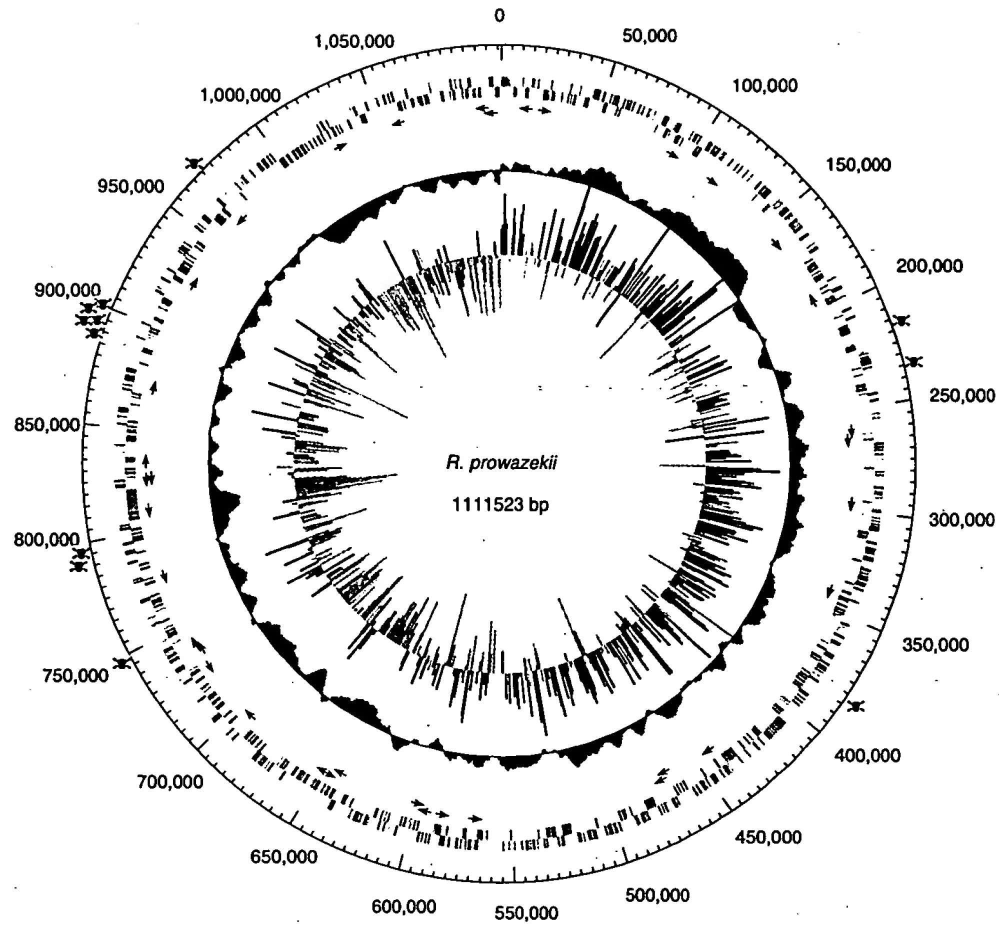


Figure 1 Overall structure of the *R. prowazekii* genome. The putative origin of replication is at 0 kb. The outer scale indicates the coordinates (in base pairs). The positions of pseudogenes are highlighted with death's heads. The distribution of genes is shown on the first two rings within the scale. The location and direction of transcription of rRNA are shown by pink arrows and of tRNA genes by black arrows. The next circle in shows GC-skew values measured over all bases in the genome. Red and purple colours denote positive and negative signs, respectively. The window size was 10,000 nucleotides and the step size was 1,000 nucleotides. The central circles shows GC-skew values calculated for third positions in the codon only. GC-skew values were calculated separately for genes located on the outer strand (green) and on the inner strand (blue). To allow easier visual inspection, the signs of the values calculated for genes located on the inner strand have been reversed.

pseudogenes (0.9% of the genome) and less than 0.2% of the genome is accounted for by non-coding repeats. The remaining 22.9% contains no open reading frames of significant length and it has the low G+C content (mean 23.7%) that is characteristic of spacer sequences in the *R. prowazekii* genome[14]. A region of 30 kilobases (kb) located at position 886–916 kb contains as much as 41.6% non-coding DNA and 11.5% pseudogenes. The non-coding DNA in this region has a small, but significantly higher, G+C content (mean 27.3%) than non-coding DNA in other areas of the genome (mean 23.7%) ($P < 0.001$), indicating that it may correspond to inactivated genes that are being degraded by mutation (Fig. 3).

Origin of replication. The origin of replication has not been experimentally identified in the *R. prowazekii* genome, but we identified *dnaA* at ~750 kb. However, the genes flanking the *dnaA* gene differ from the conserved motifs found in *Escherichia coli* and *Bacillus subtilis* (*rnpA–rpmH–dnaA–dnaN–recF–gyrB*). In *R. prowazekii*, the genes *rnpA* and *rpmH* are located in the vicinity of *dnaA*, but in the reverse orientation compared to the consensus motif, and *dnaN*, *recF* and *gyrB* are located elsewhere.

The origin and end replication in microbial genomes are often associated with transitions in GC skew (G − C/G + C) values[24]. In *R. prowazekii* we observe transitions in the GC skew values at around 0 and 500–600 kb (Fig. 1). There is a weak asymmetry in the distribution of genes in the two strands, such that the first half of the genome has a 1.6-fold higher gene density on one strand and the second half of the genome has a 1.6-fold higher gene density on the other strand. The shift in coding-strand bias correlates with the shift in GC-skew values. As most genes are transcribed in the direction of replication in microbial genomes, the origin of replication may correspond to the shift in GC-skew values at the position that we have chosen as the start point for numbering. Indeed, several short sequence stretches that are characteristic of *dnaA*-binding motifs are found in the intergenic region of genes *RP001* and *RP885* at 0 kb, supporting this interpretation.

Stable RNA sequences and repeat elements. We identified 33 genes encoding transfer RNA, corresponding to 32 different isoacceptor tRNA species. There is a single copy of each of the rRNA genes, with *rrs* located more than 500 kb away from the *rrl–rrf* gene cluster

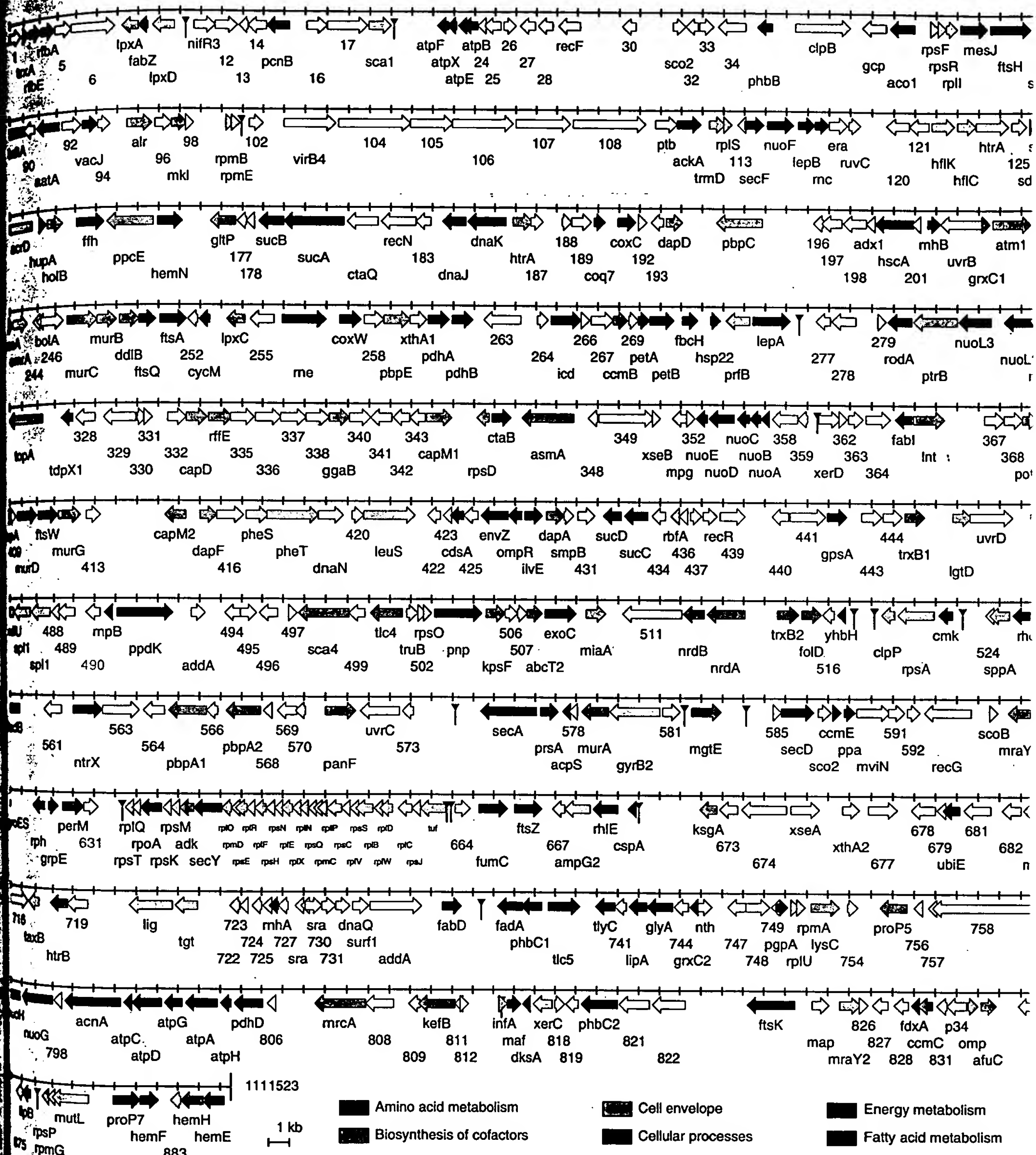
Figure 2 Linear map of the *R. prowazekii* chromosome. The position and orientation of known genes are indicated by arrows. Coding regions are colour coded according to their functional roles. The positions of tRNA genes are indicated (inverted triangle on stalk). For additional information, see http://evolution.bmc.uu.se/~siv/gnomics/Rickettsia.html.

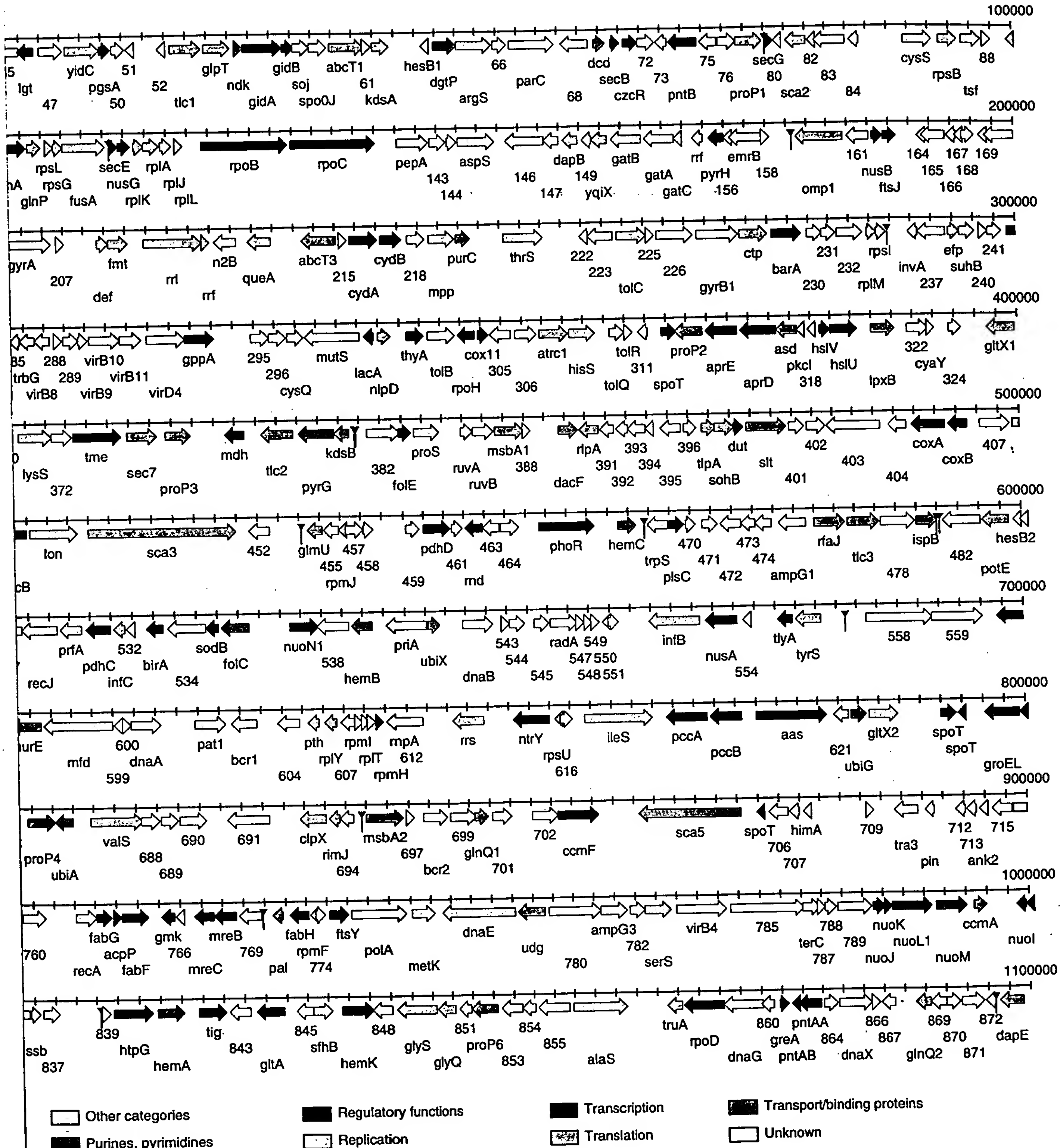
the step size was 1,000
ues calculated for third
ated separately for genes
nd (blue). To allow easier
enes located on the inner

weak asymmetry in
h that the first half of
n one strand and the
r gene density on the
rrelates with the shift
ed in the direction of
of replication may
the position that we
Indeed, several short
A-binding motifs are
1 and *RP885* at 0 kb.

Ve identified 33 genes
lifferent isoacceptor
the rRNA genes, with
rrl–rrf gene cluster.

some. The position and
oding regions are colour
ions of tRNA genes are
information, see http://





100000
5 yidC 51 glpT gidB abcT1 hesB1 66 dcd 72 75 secG 82 cysS 88
lgt pgsA 52 ndk soj 61 dgtP parC secB 73 76 80 83 rpsB
47 50 tlc1 gidA spo0J kdsA argS 68 czcR pntB proP1 sca2 84 tsf
200000
rpsL secE rplA rpoB rpoC pepA aspS dapB gatB rrf emrB 161 164 167 169
hA rpsG nusG rplJ 143 146 149 gatA pyrH 158 nusB 165 168
glnP fusA rplK rplL 144 147 yqiX gatC 156 omp1 ftsJ 166
300000
gyrA fmt n2B abcT3 cydB purC thrS 222 225 ctp 231 rpsI efp 241
207 rrl queA 215 218 223 226 barA 232 invA suhB
def rrf cydA mpp tolC gyrB1 230 rplM 237 240
400000
85 288 virB10 gppA 295 mutS thyA cox11 atrc1 tolR proP2 asd hslV 322 gltX1
trbG 289 virB11 296 lacA tolB 305 hisS 311 aprE pkcl hslU cyaY
virB8 virB9 virD4 cysQ nlpD rpoH 306 tolQ spoT aprD 318 lpxB 324
500000
0 tme mdh kdsB proS msbA1 rlpA 393 396 dut 402 coxA 407
lysS sec7 tlc2 382 ruvA 388 391 394 tlpA slt 403 coxB
372 proP3 pyrG folE ruvB dacF 392 395 sohB 401 404
600000
lon sca3 452 glmU 457 pdhD 463 phoR hemC 470 473 rfaJ ispB hesB2
455 458 461 464 trpS 471 474 tlc3 482 potE
cB rpmJ 459 rnd plsC 472 ampG1 478
700000
prfA 532 sodB nuoN1 priA 543 radA 549 infB tlyA 558 559
pdhC birA folC 538 ubiX 544 547 550 nusA tyrS
recJ infC 534 hemB dnaB 545 548 551 554
800000
urE 600 pat1 pth rpmI rnpA rrs ntrY ileS pccA aas gltX2 spoT
mfd dnaA bcr1 rplY rplT 612 rpsU pccB 621 spoT
599 604 607 rpmH 616 ubiG groEL
900000
valS 690 691 clpX msbA2 699 702 sca5 spoT himA 709 712 715
proP4 688 rimJ 697 glnQ1 ccmF 706 tra3 713
ubiA 689 694 bcr2 701 707 pin ank2
1000000
fabG gmk mreB fabH ftsY dnaE ampG3 virB4 785 788 nuoK ccmA
760 acpP 766 769 rpmF polA udg 782 terC 789 nuoL1 nuoI
recA fabF mreC pal 774 metK 780 serS 787 nuoJ nuoM
1100000
839 tig 845 848 851 854 truA 860 pntAA 866 869 872
ssb htpG 843 sfhB glyS proP6 855 rpoD greA 864 867 870 dapE
837 hemA gltA hemK glyQ 853 alaS dnaG pntAB dnaX glnQ2 871
Other categories
Purines, pyrimidines
Regulatory functions
Replication
Transcription
Translation
Transport/binding proteins
Unknown

Table 1. Functional classification of *Rickettsia prowazekii* protein-coding genes. Gene numbers correspond to those in Fig. 2. Percentages represent per cent identitites.

AMINO ACID METABOLISM 6

Amino acid biosynthesis
RP743 *glyA* serine hydroxymethyl transferase (B-*Mex*, 60.9%)
RP428 *ilvE* branched-chain amino acid aminotransferase (B-*Eco*, 36.8%)

Amino acid degradation
RP091 *aatA* aspartate aminotransferase (B-*Rme*, 55.6%)
RP618 *pccA* propionyl-CoA carboxylase α chain (E-*Rno*, 45.0%)
RP619 *pccB* propionyl-CoA carboxylase β chain (E-*Hsa*, 63.3%)
RP449 *tdcB* threonine dehydratase (E-Yeast, 35.3%)

BIOSYNTHESIS OF COFACTORS 25

Folic acid
RP536 *folC* folylpolyglutamate synthetase (B-*Bsu*, 34.5%)
RP615 *folD* methylenetetrahydrofolate dehydrogenase (*Bsu*, 46.3%)
RP383 *folE* GTP cyclohydrolase I (B-*Syn*, 48.1%)

Haem and porphyrins
RP841 *hemA* delta-aminolevulinate synthase (B-*Bja*, 49.1%)
RP539 *hemB* delta-aminolevulinate dehydratase (B-*Bja*, 53.3%)
RP466 *hemC* prophobilinogen deaminase (B-*Pmi*, 39.1%)
RP885 *hemE* uroporhyrinogen decarboxylase (B-*Rca*, 42.6%)
RP882 *hemF* coproporhyrinogen III oxidase (E-*Gma*, 43.0%)
RP884 *hemH* ferrochelatase (B-*Syn*, 40.4%)
RP847 *hemK* protoporphyrinogen oxidase (B-*Eco*, 44.3%)
RP175 *hemN* oxygen-independent coproporphyrinogen II (B-*Bsu*, 34.4%)

Lipoate
RP742 *lipA* lipoic acid synthetase (B-*Hin*, 50.5%)
RP676 *lipB* lipoic acid ligase (B-*Mtu*, 35.6%)

Menaquinone and ubiquinonens
RP190 *coq7* ubiquinone biosynthesis prt Coq7 (E-*Rno*, 36.9%)
RP479 *ispB* octaprenyl-diphosphate synthase (B-*Eco*, 36.5%)
RP696 *ubiA* 4-hydroxybenzoate octaprenyltransferase (B-*Eco*, 36.1%)
RP541 *ubiX* 3-octaprenyl-4-hydroxybenzoate carboxylase (B-*Eco*, 53.2%)
RP680 *ubiE* ubiquinone biosynthesis methyltransferase (B-*Eco*, 44.0%)
RP622 *ubiG* 3-demethylubiquinone methyltransferase (B-*Eco*, 39.1%)

Thio- and glutaredoxin
RP204 *grxC1* glutaredoxin 3 (B-*Eco*, 50.0%)
RP745 *grxC2* glutaredoxin 3 (B-*Syn*, 50.0%)
RP327 *tdpX1* thioredoxin-peroxidase (B-*Hpy*, 54.0%)
RP002 *trxA* thioredoxin (B-*Anl*, 52.8%)
RP445 *trxB1* thioredoxin reductase (B-*Hin*, 52.0%)
RP514 *trxB2* thioredoxin reductase (B-*Cpa*, 28.4%)

CELL ENVELOPE 59

Diaminopimelate
RP316 *asd* aspartate-semialdehyde dehydrogenase (B-*Vch*, 43.3%)
RP429 *dapA* dihydrodipicolinate synthase (A-*Mja*, 39.6%)
RP148 *dapB* dihydrodipicolinate reductase (B-*Hpy*, 37.7%)
RP194 *dapD* tetrahydrodipicolinate N-succinyltransferase (B-*Eco*, 57.9%)
RP874 *dapE* succinyl-diaminopimelate desuccinylase (B-*Hin*, 37.5%)
RP415 *dapF* diaminopimelate epimerase (B-*Hin*, 35.0%)
RP753 *lysC* aspartokinase (B-*Bst*, 37.3%)

Membranes and lipoproteins
RP347 *asmA* outer membrane assembly protein (B-*Eco*, 19.3%)
RP446 *lgtD* prolipoprotein diacylglyceryl transferase (B-*Vch*, 29.1%)
RP366 *lnt* apolipoprotein N-acyltransferase (B-*Hin*, 29.1%)
RP300 *nlpD* lipoprotein (B-*Hin*, 22.4%)
RP390 *rlpA* rare lipoprotein A (B-*Hin*, 23.9%)
RP224 *tolC* outer membrane protein (B-*Eco*, 22.9%)
RP048 *yidC* inner membrane protein, 60 kDa (B-*Hin*, 30.4%)

Murein sacculus
RP095 *alr* alanine racemace (B-*Hin*, 29.5%)
RP389 *dacF* penicillin binding protein precursor (B-*Bsu*, 33.8%)
RP249 *ddlB* D-alanine:D-alanine ligase (B-*Hin*, 32.8%)
RP454 *glmU* UDP-N-acetylglucoseamine pyrophosphorylase (B-*Hin*, 34.3%)
RP595 *mraY1* phospho-N-acetylmuramoyl-pentapeptide-transferase (B-*Hin*, 49.9%)
RP825 *mraY2* phospho-N-acetylmuramoyl-pentapeptide-transferase (B-*Sac*, 22.0%)
RP807 *mrcA* penicillin binding protein 1A (B-*Eco*, 35.6%)
RP579 *murA* UDP-N-acetylglucosamine 1-carboxyvinyltransferase (B-*Aca*, 51.6%)
RP248 *murB* UDP-N-acetylenolyruvoylglucosamine reductase (B-*Bsu*, 35.7%)
RP247 *murC* UDP-N-acetylmuramoylalanine ligase (B-*Hin*, 41.5%)
RP410 *murD* UDP-N-acetylmuramoylalanine-D-glutamate ligase (B-*Hin*, 32.9%)
RP597 *murE* UPD-MurNac-tripeptide synthetase (B-*Bsu*, 35.2%)
RP596 *murF* UDP-MurNac-pentapeptide synthetase (B-*Eco*, 30.6%)
RP412 *murG* UDP-MurNac-pentapeptide transferase (B-*Bsu*, 28.8%)
RP565 *pbpA1* penicillin binding protein (B-*Hin*, 34.3%)
RP567 *pbpA2* penicillin binding protein (B-*Bsu*, 30.7%)
RP195 *pbpC** penicillin-binding protein (B-*Eco*, 26.7%)
RP259 *pbpE* penicillin binding protein (B-*Bsu*, 25.2%)
RP400 *slt* lytic murein transglycosidase (B-*Hin*, 21.4%)

Surface polysaccharides, lipopolysaccharides and antigens
RP333 *capD* capsular polysaccharide biosynthesis protein CapD (B-*Sau*, 34.6%)
RP344 *capM1* capsular polysaccharide biosynthesis protein CapM (B-*Sau*, 24.9%)
RP414 *capM2* capsular polysaccharide biosynthesis protein CapM prt (B-*Sau*, 23.7%)
RP339 *ggaB* galactosamine-containing techoic acid biosynthesis (B-*Bsu*, 23.3%)
RP718 *htrB* lauroyl acetyltransferase (B-*Hin*, 21.5%)
RP007 *lpxA* UDP-N-acetylglucosamine acyltransferase (B-*Rri*, 90.4%)
RP321 *lpxB* lipid A disaccharide synthetase (B-*Hin*, 27.3%)
RP254 *lpxC* UDP-3-O-acyl N-acetylglucosamine deacetylase (B-*Eco*, 44.4%)
RP009 *lpxD* UDP-3-O-(R-3-hydroxymyristoyl)-glucosamine N-acetyltransferase (B-*Rri*, 92.4%)
RP062 *kdsA* 3-deoxy-d-manno-octulosonic acid 8-phosphate synthetase (B-*Pha*, 45.1%)
RP379 *kdsB* CTP:CMP-3-deoxy-manno-octulosonate-cytidylyl transferase (B-*Hin*, 34.8%)
RP089 *kdtA* 3-deoxy-D-manno-octulosonic acid transferase (B-*Eco*, 28.9%)
RP505 *kpsF* polysilac acid capsule expression protein (B-*Eco*, 36.9%)
RP833 *omp* cell surface antigen, 17 kD (B-*Rty*, 46.9%)
RP160 *omp1* OMP1 precursor (B-*Bab*, 29.5%)
RP771 *pal* peptidoglycan-associated lipoprotein (B-*Eco*, 37.9%)
RP476 *rfaJ* lipopolysaccharide 1,2-glucosyltransferase (B-*Sty*, 26.1%)
RP004 *rfbA* O-antigen export system permease (B-*Kpn*, 22.3%)
RP003 *rfbE* O-antigen ABC export system, ATP-binding protein (B-*Yen*, 34.0%)
RP334 *rffE* UDP-N-acetylglucosamine 2-epimerase (B-*Bsu*, 26.8%)
RP018 *sca1* cell surface antigen (B-*Rri*, 24.9%)
RP081 *sca2* cell surface antigen (B-*Rri*, 27.4%)
RP451 *sca3* cell surface antigen (B-*Rri*, 27.6%)
RP498 *sca4* cell surface antigen (B-*Rja*, 67.4%)
RP704 *sca5* cell surface antigen (B-*Rri*, 72.5%)
RP779 *udg* UDP-glucose 6-dehydrogenase (B-*Pae*, 31.8%)

CELLULAR PROCESSES 44

Cell division
RP251 *ftsA* cell division protein FtsA (B-*Hin*, 29.5%)
RP043 *ftsH* cell division protein FtsH (B-*Eco*, 54.0%)
RP163 *ftsJ* cell division protein FtsJ (B-*Eco*, 44.4%)
RP823 *ftsK* cell division protein FtsK (B-*Cbu*, 41.5%)
RP250 *ftsQ* cell division protein FtsQ (B-*Hin*, 17.9%)
RP411 *ftsW* cell division protein FtsW (B-*Eco*, 33.2%)
RP775 *ftsY* cell division protein FtsY (B-*Hin*, 43.2%)
RP666 *ftsZ* cell division protein FtsZ (B-*Wsp*, 65.3%)
RP056 *gidA* glucose inhibited division protein A (B-*Eco*, 48.8%)
RP057 *gidB* glucose inhibited division protein (B-*Ppu*, 26.8%)
RP815 *maf* MAF protein (B-*Bsu*, 38.1%)
RP042 *mesJ* cell cycle protein MesJ (B-*Eco*, 22.1%)
RP768 *mreB* rod shape-determining protein (B-*Eco*, 60.5%)
RP767 *mreC* rod shape-determining protein (B-*Eco*, 23.1%)
RP280 *rodA* rod shape-determining protein (B-*Eco*, 38.1%)

Cell killing
RP555 *tlyA* hemolysin (B-*Thy*, 34.3%)
RP740 *tlyC* hemolysin (B-*Thy*, 28.8%)

Chaperones and stress-induced proteins
RP670 *cspA* cold shock protein (B-*Scl*, 57.6%)
RP816 *dksA* DnaK suppressor protein (B-*Hin*, 38.8%)
RP184 *dnaJ* heat shock protein (B-*Bov*, 49.7%)
RP185 *dnaK* heat shock protein 70 (B-*Rme*, 72.7%)
RP626 *groEL* heat shock protein GroEL (B-*Rme*, 69.4%)
RP627 *groES* heat shock protein GroES (B-*Rty*, 98.9%)
RP629 *grpE* heat shock protein GrpE (B-*Ccr*, 39.5%)
RP200 *hscA* heat shock protein A (B-*Hin*, 39.6%)
RP320 *hslU* heat shock protein HslU (B-*Bsu*, 54.8%)
RP319 *hslV* heat shock prt HslV (B-*Hin*, 54.1%)
RP273 *hsp22* heat shock protein (E-*Phy* cp, 29.0%)
RP840 *htpG* heat shock protein C62.5 (B-*Eco*, 43.1%)

Detoxification
RP535 *sodB* superoxide dismutase (B-*Lpn*, 53.4%)
RP759 *thdF* thiophene and furan oxidizer (B-*Hin*, 34.7%)

Protein and peptide secretion
RP315 *aprD* protease secretion ATP-binding protein (B-*Pae*, 40.0%)
RP314 *aprE* protease secretion ATP-binding protein (B-*Pae*, 32.4%)
RP173 *ffh* signal recognition particle receptor protein (B-*Eco*, 49.6%)
RP275 *lepA* GTP-binding membrane protein (B-*Bsu*, 57.0%)
RP116 *lepB* signal peptidase (B-*Sty*, 37.3%)
RP575 *secA* preprotein translocase SecA subunit (B-*Rca*, 51.8%)
RP070 *secB* preprotein translocase SecB subunit (B-*Hin*, 30.7%)
RP586 *secD* protein-export membrane protein (B-*Eco*, 40.4%)
RP134 *secE* preprotein translocase SecE subunit (B-*Bsu*, 37.3%)
RP114 *secF* protein-export membrane protein (B-*Hin*, 37.7%)
RP079 *secG* protein-export membrane protein (B-*Hpy*, 32.0%)
RP639 *secY* preprotein translocase SecY subunit (B-*Eco*, 50.0%)
RP842 *tig* trigger factor (B-*Eco*, 32.0%)

ENERGY METABOLISM 67

ATP-proton motive force interconversion
RP803 *atpA* ATP synthase F1 alpha subunit (B-*Rru*, 66.2%)
RP023 *atpB* ATP synthase F0 subunit a (B-*Rru*, 51.5%)
RP800 *atpC* ATP synthase F1 epsilon subunit (B-*Rru*, 24.5%)
RP801 *atpD* ATP synthase F1 beta subunit (B-*Rca*, 77.0%)
RP022 *atpE* ATP synthase F0 subunit c (E-*Ram* mt, 43.2%)
RP020 *atpF* ATP synthase F0 subunit b (B-*Rru*, 21.1%)
RP802 *atpG* ATP synthase F1 gamma subunit (B-*Rbl*, 38.0%)
RP804 *atpH* ATP synthase F1 delta chain (E-*Osl* cp 26.4%)
RP021 *atpX* ATP synthase F0 subunit b' (E-*Ram* mt, 28.6%)

Electron transport
RP588 *ccmE* cytochrome *c* biogenesis protein (B-*Hin*, 33.2%)
RP703 *ccmF* cytochrome *c* biogenesis protein (E-*Ram* mt, 33.6%)
RP405 *coxA* cytochrome *c* oxidase subunit I (E-*Mpo* mt, 68.9%)
RP406 *coxB* cytochrome *c* oxidase subunit II (E-*Mpo* mt, 48.0%)
RP191 *coxC* cytochrome *c* oxidase subunit III (E-*Pwi* mt, 48.9%)
RP257 *coxW* cytochrome *c* oxidase assemly (E-*Sce*, 35.2%)
RP304 *cox11* cytochrome *c* oxidase assembly protein (E-*Ram* mt, 42.9%)
RP346 *ctaB* cytochrome *c* oxidase assembly factor (B-*Pde*, 40.6%)
RP253 *cycM* cytochrome c (B-*Bja*, 35.6%)
RP216 *cydA* cytochrome oxidase *d* subunit I (B-*Avi*, 34.0%)
RP217 *cydB* cytochrome oxidase *d* subunit II (B-*Eco*, 30.0%)
RP272 *fbcH* ubiquinol cytochrome *c* oxidoreductase, cytochrome c1 subunit (B-*Bja*, 47.8%)
RP829 *fdxA* ferredoxin (*Rca*, 57.5%)
RP357 *nuoA* NADH dehydrogenase I chain A (E-*Pwi* mt, 64.6%)
RP356 *nuoB* NADH dehydrogenase I chain B (E-*Ram* mt, 73.2%)
RP355 *nuoC* NADH dehydrogenase I chain C (B-*Pde*, 42.1%)
RP354 *nuoD* NADH dehydrogenase I chain D (E-*Ram* mt, 71.4%)
RP353 *nuoE* NADH dehydrogenase I chain E (E-*Rno*, mt, 55.1%)
RP115 *nuoF* NADH dehydrogenase I chain F (B-*Pde*, 69.1%)
RP797 *nuoG* NADH dehydrogenase I chain G (E-*Bta* mt, 49.3%)
RP796 *nuoH* NADH dehydrogenase I chain H (E-*Ram* mt, 63.5%)
RP795 *nuoI* NADH dehydrogenase I chain I (E-*Ram* mt, 71.4%)
RP794 *nuoJ* NADH dehydrogenase I chain J (E-*Ram* mt 42.3%)
RP791 *nuoK* NADH dehydrogenase I chain K (B-*Pde*, 61.4%)
RP792 *nuoL1* NADH dehydrogenase I chain L (E-*Ram* mt, 45.5%)
RP283 *nuoL2* NADH dehydrogenase I chain L (E-*Aja* cp, 27.0%)
RP282 *nuoL3* NADH dehydrogenase I chain L (E-*Ame* mt, 17.0%)
RP793 *nuoM* NADH dehydrogenase I chain M (E-*Ram* mt, 46.6%)
RP537 *nuoN1* NADH dehydrogenase I chain N (E-*Ram* mt, 34.6%)
RP284 *nuoN2* NADH dehydrogenase I chain N (E-*Ram* mt, 23.2%)
RP270 *petA* Rieske-I iron sulphur protein (B-*Bja*, 58.3%)
RP271 *petB* cytochrome *b* (B-*Rru*, 65.4%)
RP863 *pntAA* NAD(P) transhydrogenase α subunit (B-*Eco*, 37.7%)
RP862 *pntAB* NAD(P) transhydrogenase α subunit (B-*Hin*, 44.7%)
RP074 *pntB* NAD(P) transhydrogenase β subunit (B-*Hin*, 51.5%)

Fermentation
RP110 *ackA* acetate kinase (B-*Cts*, 38.4%)

Glycolysis
RP492 *ppdK* pyruvate, orthophosphate dikinase (E-*Fir*, 48.8%)

Phosphate
RP589 *ppa* inorganic pyrophosphatase (B-*Eco*, 59.3%)

Pyruvate dehydrogenase
RP261 *pdhA* pyruvate dehydrogenase E1 component, α subunit (E-*Ath*, 44.0%)
RP262 *pdhB* pyruvate dehydrogenase E1 component, β subunit (E-*Sce*, 59.7%)
RP530 *pdhC* dihydrolipoamide acetyltransferase E2 component (E-*Rno*, 45.1%)
RP460 *pdhD* dihydrolipoamide dehydrogenase E3 component (E-*Psa*, 54.7%)
RP805 *pdhD* dihydrolipoamide dehydrogenase E3 component (*Zym*, 51.1%)

TCA cycle
RP799 *acnA* aconitate hydratase (B-*Lpn*, 59.1%)
RP665 *fumC* fumarate hydratase (B-*Ror*, 63.5%)
RP844 *gltA* citrate synthase (B-*Rty*, 97.8%)
RP265 *icd* isocitrate dehydrogenase (B-*Tth*, 38.6%)
RP376 *mdh* malate dehydrogenase (B-*Can*, 51.5%)
RP128 *sdhA* succinate dehydrogenase, flavoprotein subunit (B-*Bja*, 70.0%)
RP044 *sdhB* succinate dehydrogenase, iron-sulphur protein (E-*Ram* mt, 69.0%)
RP126 *sdhC* succinate dehydrogenase, cytochrome b_{556} subunit (E-*Ram* mt, 39.5%)
RP127 *sdhD* succinate dehydrogenase, subunit IV (E-*Ram* mt, 25.6%)
RP180 *sucA* 2-oxoglutarate dehydrogenase (B-*Hin*, 44.3%)
RP179 *sucB* dihydrolipoamide succinyltransferase (B-*Eco*, 48.7%)
RP433 *sucC* succinyl-CoA synthetase, β subunit (B-*Eco*, 52.1%)
RP432 *sucD* succinyl-CoA synthetase, α subunit (E-*Ddi*, 70.7%)

Sugars
RP509 *exoC* phosphomannomutase (B-*Abr*, 42.7%)
RP299 *lacA* galactosidase acetyltransferase (B-*Mpn*, 44.4%)

FATTY ACID AND PHOSPHOLIPID METABOLISM 25

RP620 *aas* 2-acyl-glycerol-phosphate-ethanolamine (B-*Eco*, 39.9%)
RP038 *aco1* acyl-CoA desaturase (E-Yeast, 27.6%)
RP763 *acpP* acyl carrier protein (B-*Lmu*, 52.6%)
RP577 *acpS* holo-[acyl carrier protein] synthase (B-*Eco*, 38.5%)
RP533 *birA* biotin Ac-CoA carboxylase synthase (B-*Pde*, 33.6%)
RP424 *cdsA* phosphatidate cytidylyltransferase (B-*Eco*, 31.5%)
RP735 *fabD* malonyl-CoA:Acyl carrier protein transacylase (B-*Bsu*, 40.3%)
RP764 *fabF* 3-oxoacyl-[acyl-carrier-protein] synthase II(B-*Eco*, 53.5%)
RP762 *fabG* 3-oxoacyl-[acyl-carrier-protein] reductase (B-*Rme*, 54.8%)
RP772 *fabH* 3-oxoacyl-[acyl-carrier-protein] synthase (B-*Rca*, 47.3%)
RP365 *fabI* enoyl-[acyl-carrier-protein] reductase (B-*Asp*, 49.0%)
RP008 *fabZ* 3R-hydroxymyristoyl acyl carrier protein dehydratase (B-*Rri*, 91.7%)
RP737 *fadA* acetyl-CoA acetyltransferase (B-*Pde*, 54.5%)
RP560 *fadB* fatty oxidation complex a subunit (E-*Cel*, 30.6%)
RP442 *gpsA* glycerol-3-phosphate dehydrogenase (B-*Eco*, 32.1%)
RP046 *lgt* prolipoprotein diacylglycerol (B-*Eco*, 39.1%)
RP035 *phbB* acetoacetyl-CoA reductase (B-*Zra*, 52.9%)
RP738 *phbC1* poly-beta-hydroxybutyrate polymerase (B-*Cvi*, 22.7%)
RP820 *phbC2* poly-beta-hydroxybutyrate polymerase (B-*Max*, 37.4%)
RP750 *pgpA* phosphatidylglycerophosphatase A (B-*Hin*, 31.9%)
RP049 *pgsA* phospatidylglycerophosphate synthase (B-*Bsu*, 40.1%)
RP469 *plsC* 1-acyl-glycerol-3-phosphate acyltransferase (E-*Sce*, 23.6%)
RP242 *pssA* phosphatidylserine synthase (B-*Hpy*, 28.8%)
RP373 *tme* malic enzyme (B-*Hin*, 45.5%)
RP093 *vacJ* VacJ lipoprotein precursor (B-*Sfl*, 33.8%)

PURINES, PYRIMIDINES 14

Deoxyribonucleotide metabolism
RP169 *dcd* deoxycytidine triphosphate deaminase (B-*Hin*, 27.8%)
RP064 *dgtP* deoxyguanosine triphosphate triphosphohydrolase (B-*Eco*, 30.1%)
RP399 *dut* deoxyuridine 5' triphosphate nucleotidohydrolase (B-*Hin*, 43.1%)
RP055 *ndk* nucleoside diphosphate kinase (B-*Rsu*, 56.4%)
RP513 *nrdA* ribonucleotide redutase α chain (E-*Sce*, 33.7%)
RP512 *nrdB* ribonucleotide reductase β chain (E-*Mmu*, 28.5%)
RP301 *thyA* thymidylate synthase (E-*Ddi* 47.9%)
RP684 *tmk* thymidylate kinase (B-*Bsu*, 38.7%)

Purine ribonucleotide metabolism
RP638 *adk* adenylate kinase (B-*Sco*, 33.8%)
RP765 *gmk* guanylate kinase (B-*Hin*, 49.0%)
RP220 *purC* phosphoribosylaminoimidazole-succinocarboxamide synthase (B-*Spn*, 34.0%)

Pyrimidine ribonucleotide biosynthesis
RP522 *cmk* cytidylate kinase (B-*Eco*, 34.9%)
RP378 *pyrG* CTP synthase (B-*Abr*, 54.7%)

step size was 1,000
s calculated for third
d separately for genes
(blue). To allow easier
es located on the inner

eak asymmetry in
that the first half of
one strand and the
gene density on the
elates with the shift
in the direction of
of replication may
e position that we
deed, several short
binding motifs are
and *RP885* at 0 kb

identified 33 genes
ferent isoacceptor
e rRNA genes, with
rl–rrf gene cluster

ne. The position and
ng regions are colour
s of tRNA genes are
formation, see http:

RP155 pyrH uridylate kinase (B-Syn, 53.3%)

REGULATORY FUNCTIONS 14

RP229 barA histidine kinase sensor protein (B-Eco, 23.2%)
RP071 czcR transcriptional activator protein (B-Aeu, 35.1%)
RP426 envZ histidine kinase osmolarity sensor protein (B-Stu, 23.65)
RP294 gppA pppGpp phosphohydrolase (B-Hpy, 23.3%)
RP011 nifR3 transcriptional activator nitrogen assimilation protein (B-Abr, 59.0%)
RP614 ntrY histidine kinase nitrogen sensor protein (B-Aca, 30.6%)
RP562 ntrX transcriptional activator nitrogen assimilation protein (B-Aca, 45.2%)
RP427 ompR transcriptional activator protein OmpR (B-Rca, 42.9%)
RP465 phoR histidine kinase phosphatase synthesis sensor protein (B-Bsu, 24.4%)
RP312 spoT* (p)ppGpp 3'-pyrophosphohydrolase (B-Eco, 29.9%)
RP624 spoT* (p)ppGpp 3'-pyrophosphohydrolase (B-Mpn, 27.8%)
RP625 spoT* (p)ppGpp 3'-pyrophosphohydrolase (B-Eco, 48.7%)
RP705 spoT* (p)ppGpp 3'-pyrophosphohydrolase (B-Eco, 31.7%)
RP517 yhbH sigma 54 modulation protein (B-Bja, 26.2%)

REPLICATION 46

Degradation of DNA

RP734 addA ATP-dependent nuclease (B-Bsu, 23.7%)
RP260 xthA1 exodeoxyribonuclease III (B-Eco, 30.1%)
RP676 xthA2 exodeoxyribonuclease III (B-Eco, 33.2%)
RP675 xseA exodeoxyribonuclease large subunit (B-Eco, 31.7%)
RP350 xseB exodeoxyribonuclease small subunit (B-Eco, 32.5%)

DNA replication, restriction, modification, recombination and repair

RP601 dnaA chromosomal replication initiation protein DnaA (B-Eco, 44.1%)
RP542 dnaB DNA helicase (E-Osl cp, 40.9%)
RP778 dnaE DNA polymerase III alpha subunit (B-Sty, 37.2%)
RP859 dnaG DNA primase (B-Sty, 29.0%)
RP419 dnaN DNA polymerase III beta subunit (B-Ppu, 29.9%)
RP732 dnaQ DNA polymerase III epsilon subunit (B-Sty, 46.7%)
RP865 dnaX DNA polymerase III gamma chain (B-Eco, 31.4%)
RP206 gyrA DNA gyrase A subunit (B-Rsp, 49.4%)
RP227 gyrB1 DNA gyrase B subunit (B-Scl, 42.0%)
RP580 gyrB2 DNA gyrase B subunit (B-Ppu, 51.5%)
RP172 holB DNA polymerase III, delta prime subunit (B-Pae, 22.3%)
RP171 hupA DNA binding protein HU (B-Vpr, 47.8%)
RP720 lig DNA ligase (B-Zmo, 45.7%)
RP777 metK* S-adenosylmethionine synthetase (B-Eco, 66.3%)
RP598 mfd transcription-repair coupling factor (B-Hin, 33.9%)
RP351 mpg DNA-3-methyladenine glycosidase (E-Hsa, 29.7%)
RP880 mutL DNA mismatch repair protein MutL (B-Spn, 35.4%)
RP298 mutS DNA mismatch repair protein MutS (B-Bsu, 39.0%)
RP746 nth endonuclease III (B-Eco, 50.7%)
RP067 parC DNA topoisomerase IV subunit A (B-Hin, 39.0%)
RP711 pin* invertase/recombinase (B-Eco, 38.0%)
RP776 polA DNA polymerase I (B-Bca, 37.2%)
RP540 priA primosomal protein replication factor (B-Rru, 39.7%)
RP546 radA DNA repair (B-Bsu, 46.5%)
RP761 recA recombination protein RecA (B-Pde, 71.2%)
RP029 recF DNA repair protein, ATP binding protein (B-Ccr, 30.4%)
RP593 recG ATP-dependent DNA helicase (B-Eco, 34.1%)
RP528 recJ single-stranded DNA-specific exonuclease (B-Eco, 32.8%)
RP182 recN recombination protein RecN (B-Hin, 31.6%)
RP438 recR recombination protein RecR (B-Bsu, 36.9%)
RP385 ruvA Holliday junction DNA helicase (B-Pae, 35.0%)
RP386 ruvB Holliday junction DNA helicase (Pae, 51.5%)
RP119 ruvC Holliday junction endodeoxyribonuclease (B-Eco, 36.1%)
RP836 ssb single-stranded binding protein (B-Bab, 52.6%)
RP326 topA DNA topoisomerase I (B-Bsu, 44.9%)
RP835 uvrA repair excision nuclease subunit A (B-Eco, 57.7%)
RP203 uvrB repair excision nuclease subunit B (B-Hin, 56.0%)
RP572 uvrC repair excision nuclease subunit C (B-Pfl, 36.9%)
RP447 uvrD DNA helicase (B-Sau, 43.5%)
RP817 xerC integrase/recombinase (B-Bsu, 32.2%)
RP361 xerD integrase/recombinase (B-Eco, 37.6%)

TRANSCRIPTION 20

Degradation of RNA

RP504 pnp polyribonucleotide nucleotidyltransferase (B-Eco, 48.9%)
RP117 rnc ribonuclease III (B-Hpy, 40.2%)
RP462 rnd ribonuclease D (B-Eco, 28.5%)
RP256 rne ribonuclease E (B-Eco, 35.9%)
RP726 rnhA ribonuclease HI (B-Msm, 43.4%)
RP202 rnhB ribonuclease HII (B-Eco, 44.7%)
RP611 rnpA ribonuclease P (B-Mca, 28.4%)
RP628 rph ribonuclease PH (B-Hin, 55.05%)

RNA synthesis and modification

RP861 greA transcription elongation factor GreA (B-Hin, 61.4%)
RP553 nusA transcription termination factor NusA (B-Eco, 36.9%)
RP162 nusB transcription termination factor NusB (B-Eco, 32.9%)
RP135 nusG transcription antitermination protein NusG (B-Eco, 42.2%)
RP015 pcnB poly (A) polymerase I (B-Bsu, 26.3%)
RP669 rhlE ATP-dependent RNA helicase (B-Eco, 38.3%)
RP526 rho transcription termination factor Rho (B-Rsp, 72.0%)
RP635 rpoA RNA polymerase alpha subunit (B-Bpe, 47.2%)
RP140 rpoB RNA polymerase beta subunit (B-Rty, 87.4%)
RP141 rpoC RNA polymerase beta' subunit (B-Eco, 58.8%)
RP303 rpoH RNA polymerase sigma-32 factor (B-Atu, 52.0%)
RP858 rpoD RNA polymerase sigma-70 factor (B-Rca, 50.5%)

TRANSLATION 118

Aminoacyl-tRNA synthetases

RP856 alaS alanyl-tRNA synthetase (B-Bba, 52.7%)
RP065 argS arginyl-tRNA synthetase (B-Hpy, 33.0%)
RP145 aspS aspartyl-tRNA synthetase (B-Syn, 43.3%)
RP085 cysS cysteinyl-tRNA synthetase (B-Hin, 46.0%)
RP325 gltX1 glutamyl-tRNA synthetase (B-Abr, 45.6%)
RP623 gltX2 glutamyl-tRNA synthetase (B-Hpy, 40.3%)
RP850 glyQ glycyl-tRNA synthetase (B-Mca, 60.6%)
RP849 glyS glycyl-tRNA synthetase (B-Bsu, 32.9%)
RP308 hisS histidyl-tRNA synthetase (B-Eco, 38.3%)
RP620 ileS isoleucyl-tRNA synthetase (B-Mtu, 48.6%)
RP421 leuS leucyl-tRNA synthetase (B-Eco, 45.3%)
RP371 lysS lysyl-tRNA synthetase (B-Bbu, 26.3%)
RP683 metS methionyl-tRNA synthetase (B-Bsu, 46.9%)
RP417 pheS phenylalanyl-tRNA synthetase alpha sub (B-Hin, 49.2%)
RP418 pheT phenylalanyl-tRNA synthetase beta sub (B-Hin, 33.9%)
RP384 proS proline-tRNA synthetase (B-Zmo, 51.8%)
RP783 serS seryl-tRNA synthetase (B-Cbu, 47.2%)
RP221 thrS threonyl-tRNA synthetase (B-Hin, 50.6%)
RP468 trpS tryptophanyl-tRNA synthetase (B-Syn, 48.5%)
RP556 tyrS tyrosyl-tRNA synthetase (B-Bca, 38.7%)
RP687 valS valyl-tRNA synthetase (A-Mja, 38.3%)

tRNA and amino acyl-tRNA modification

RP208 def methionyl-tRNA deformylase (B-Eco, 49.4%)
RP209 fmt methinoyl-tRNA formyltransferase (B-Hin, 41.9%)
RP152 gatA glutamyl-tRNA (Gln) amidotransferase subunit A (B-Mca, 48.6%)
RP151 gatB glutamyl-tRNA (Gln) amidotransferase subunit B (B-Mca, 46.9%)
RP153 gatC glutamyl-tRNA (Gln) amidotransferase subunit C (B-Bsu, 24.7%)
RP672 ksgA dimethyladenosine transferase (B-Bsu, 35.7%)
RP510 miaA tRNA delta-2-isopentenylpyrophosphate (IPP) transferase (B-Atu, 30.7%)
RP605 pth peptidyl-tRNA hydrolase (B-Hin, 40.5%)
RP213 queA S-adenosylmethionine:tRNA ribosyltransferase-isomerase (B-Hin, 43.3%)
RP721 tgt tRNA-guanine transglycosylase (B-Zmo, 61.2%)
RP111 trmD tRNA (guanine-N1)-methyltransferase (B-Eco, 44.7%)
RP857 truA pseudouridylate synthase I (B-Eco, 40.1%)
RP501 truB tRNA pseudouridine 5S synthase (B-Hin, 37.6%)

Degradation of proteins, peptides and glycopeptides

RP036 clpB ATP-dependent protease, ATP binding subunit (B-Hin, 54.3%)
RP520 clpP ATP-dependent Clp protease (B-Yen, 67%)
RP692 clpX ATP-dependent protease, ATPase subunit (B-Eco, 62.8%)
RP228 ctp tail-specific protease precursor (B-Bba, 42.6%)
RP037 gcp sialoglycoprotein endopeptidase (B-Hin, 42.2%)
RP123 hflC lambda cII stability-governing protein (B-Eco, 33.9%)
RP122 hflK lambda cII stability-governing protein (B-Vpa, 30.3%)
RP124 htrA serine protease (B-Bhe, 37.7%)
RP186 htrA protease DO (E-Sce, 26.7%)
RP450 lon ATP-dependent protease LA (B-Ccr, 53.1%)
RP408 lspA lipoprotein signal peptidase (B-Bsu, 27.9%)
RP824 map methionyl aminopeptidase (B-Sty, 55.3%)
RP219 mpp mitochondrial protease (B-Bsu, 35.4%)
RP142 pepA aminopeptidase A (B-Pae, 36.6%)
RP174 ppcE peptidase II (B-Rsn, 32.5%)
RP281 ptrB protease II (B-Eco, 34.2%)
RP398 sohB protease IV (B-Mja, 23.9%)
RP525 sppA protease IV (B-Hin, 27.6%)

Protein modification and translation factors

RP238 efp elongation factor P (B-Bsu, 39.5%)
RP132 fusA elongation factor G (B-Atu, 68.7%)
RP814 infA initiation factor IF-1 (B-Hin, 67.1%)
RP552 infB initiation factor IF-2 (B-Hin, 42.6%)
RP531 infC initiation factor IF-3 (B-Pvu, 47.7%)
RP529 prfA peptide chain release factor RF-1 (B-Bsu, 50.1%)
RP274 prfB peptide chain release factor RF-2 (B-Eco, 50.4%)
RP435 rbfA ribosome binding factor A (B-Bsu, 31.6%)
RP693 rimJ ribosome protein alanine acetyltransferase (B-Eco, 23.2%)
RP154 rrf ribosome recycling factor (B-Hin, 43.3%)
RP397 tlpA thiol:disulphide interchange protein (B-Bja, 27.4%)
RP661 tuf elongation factor Tu (B-Tcu, 81.5%)
RP087 tsf elongation factor Ts (B-Scl, 40.7%)

Ribosomal proteins; synthesis and modification

RP137 rplA ribosomal protein L1 (B-Cgr, 50.2%)
RP656 rplB ribosomal protein L2 (E-Ram mt, 61.5%)
RP659 rplC ribosomal protein L3 (E-Sce, 44.1%)
RP658 rplD ribosomal protein L4 (B-Bst, 39.3%)
RP647 rplE ribosomal protein L5 (B-Aco, 53.6%)
RP644 rplF ribosomal protein L6 (B-Bst, 45.4%)
RP041 rplI ribosomal protein L9 (E-Ppu cp, 33.6%)
RP138 rplJ ribosomal protein L10 (B-Laf, 36.7%)
RP136 rplK ribosomal protein L11 (E-Ram mt, 45.5%)
RP139 rplL ribosomal protein L7/L12 (B-Bab, 66.9%)
RP233 rplM ribosomal protein L13 (B-Sca, 52.8%)
RP649 rplN ribosomal protein L14 (B-Eco, 69.6%)
RP640 rplO ribosomal protein L15 (B-Bst, 46.5%)
RP652 rplP ribosomal protein L16 (B-Aac, 53.3%)
RP634 rplQ ribosomal protein L17 (B-Eco, 57.5%)
RP643 rplR ribosomal protein L18 (B-Bst, 58.6%)
RP112 rplS ribosomal protein L19 (B-Eco, 58.8%)
RP609 rplT ribosomal protein L20 (B-Psy, 61.5%)
RP751 rplU ribosomal protein L21 (E-Sol, 42.3%)
RP654 rplV ribosomal protein L22 (B-Eco, 50.0%)
RP657 rplW ribosomal protein L23 (B-Bst, 46.3%)
RP648 rplX ribosomal protein L24 (B-Bst, 55.9%)
RP606 rplY ribosomal protein L25 (B-Mtu, 26.9%)
RP752 rpmA ribosomal protein L27 (E-Ram mt, 62.9%)
RP099 rpmB ribosomal protein L28 (B-Mge, 43.7%)
RP651 rpmC ribosomal protein L29 (B-Bst, 39.4%)
RP641 rpmD ribosomal protein L30 (B-Mlu, 33.3%)
RP100 rpmE ribosomal protein L31 (B-Mtu, 31.6%)
RP773 rpmF ribosomal protein L32 (B-Hin, 49.1%)
RP879 rpmG ribosomal protein L33 (B-Hin, 51.8%)
RP610 rpmH ribosomal protein L34 (B-Ppu, 65.9%)
RP608 rpmI ribosomal protein L35 (E-Ppu cp, 38.5%)
RP456 rpmJ ribosomal protein L36 (E-Gth cp, 65.8%)
RP521 rpsA ribosomal protein S1 (B-Rme, 48.6%)
RP086 rpsB ribosomal protein S2 (B-Mtu, 41.5%)
RP653 rpsC ribosomal protein S3 (B-Bst, 54.2%)
RP345 rpsD ribosomal protein S4 (B-Bsu, 43.2%)
RP642 rpsE ribosomal protein S5 (B-Bst, 50.6%)
RP029 rpsF ribosomal protein S6 (B-Hin, 30.2%)
RP131 rpsG ribosomal protein S7 (E-Ram mt, 45.2%)
RP645 rpsH ribosomal protein S8 (B-Bsu, 42.0%)
RP234 rpsI ribosomal protein S9 (B-Bst, 48.8%)
RP660 rpsJ ribosomal protein S10 (B-Hin, 60.8%)
RP636 rpsK ribosomal protein S11 (B-Syn, 53.5%)
RP130 rpsL ribosomal protein S12 (B-Tth, 60.5%)
RP637 rpsM ribosomal protein S13 (B-Bst, 58.8%)
RP646 rpsN ribosomal protein S14 (B-Syn, 47.0%)
RP503 rpsO ribosomal protein S15 (B-Bst, 53.4%)
RP878 rpsP ribosomal protein S16 (B-Hin, 45.1%)
RP650 rpsQ ribosomal protein S17 (B-Tth, 61.8%)
RP040 rpsR ribosomal protein S18 (B-Syn, 50.7%)
RP655 rpsS ribosomal protein S19 (B-Eco, 58.2%)
RP03? rpsT ribosomal protein S20 (B-Rme, 40.9%)
RP615 rpsU ribosomal protein S21 (B-Bbu, 33.3%)

TRANSPORT AND BINDING PROTEINS 38

General

RP060 abcT1 ABC transporter, ATP-binding protein (B-Hin, 55.7%)
RP508 abcT2 ABC transporter, ATP-binding protein (B-Rme, 48.8%)
RP214 abcT3 ABC transporter, ATP-binding protein (B-Hin, 33.6%)
RP387 msbA1 ABC transporter, ATP-binding protein (B-Eco, 26.2%)
RP696 msbA2 ABC transporter, ATP-binding protein (B-Eco, 28.2%)

Amino acids

RP307 atrc1 cationic amino acid transporter (E-Mmu, 29.7
RP129 glnP glutamine transport system permease (B-Bsu, 48.6%)
RP700 glnQ1 glutamine ABC transporter, ATP-binding pro Eco, 39.1%)
RP868 glnQ2 glutamine ABC transporter, ATP-binding pro Bst, 51.0%)
RP176 gltP glutamate-aspartate transporter (B-Bca, 35.2
RP483 potE putrescine-ornithine transporter (B-Hin, 26.9
RP369 potG putrescine ABC transporter, ATP-binding pro Mpn, 29.2%)
RP077 proP1 proline/betain transporter (B-Eco, 26.7%)
RP313 proP2 proline/betain transporter (B-Eco, 24.9%)
RP375 proP3 proline/betain transporter (B-Eco, 21.2%)
RP685 proP4 proline/betain transporter (B-Eco, 24.0%)
RP755 proP5 proline/betain transporter (B-Eco, 27.8%)
RP852 proP6 proline/betain transporter (B-Eco, 34.8%)
RP881 proP7 proline/betain transporter (B-Eco, 28.7%)
RP150 yqiX amino acid ABC transporter (B-Bsu, 32.4%)

Nucleosides and nucleotides

RP097 mkl ribonucleotide ABC transporter, ATP-binding p (B-Mle, 36.2%)
RP053 tlc1 ATP/ADP translocase (B-Ctr, 43.3%)
RP377 tlc2 ATP/ADP translocase (B-Ctr, 35.2%)
RP477 tlc3 ATP/ADP translocase (B-Ctr, 39.6%)
RP500 tlc4 ATP/ADP translocase (B-Ctr, 36.3%)
RP739 tlc5 ATP/ADP translocase (B-Ctr, 34.7%)

Carbohydrates, organic alcohols and acids

RP054 glpT glycerol-3-phosphate permease (B-Bsu, 37.1

Cations

RP834 afuC iron ABC transporter, ATP-binding protein (B- 33.9%)
RP810 kefB glutathionine-regulated potassium-efflux sys tein (B-Eco, 33.9%)
RP583 mgtE magnesium transporter (B-Syn, 27.0%)

Other

RP205 atm1 mitochondrial ABC transporter, ATP-binding p (E-Sce, 43.3%)
RP794 ccmA haem ABC transporter A, ATP-binding protein Hin, 35.5%)
RP268 ccmB haem exporter protein B (E-Ram mt, 20.9%)
RP830 ccmC haem exporter protein C (B-Bja, 43.7%)
RP571 panF panthotenate permease (B-Hin, 20.5%)
RP630 perM permease PerM homologue (B-Hin, 25.0%)
RP374 sec7 transport protein Sec7 (B-Hsa, 29.6%)
RP576 prsA protein export (B-Bsu, 28.9%)

OTHER CATEGORIES ....

Adaptations to atypical conditions

RP708 himA integration host factor α (B-Eco, 29.5%)
RP236 invA invasion protein A (B-Bba, 42.8%)
RP590 mviN virulence factor Mvin protein (B-Sty, 32.4%)
RP717 taxB* conjugative DNA processing (B-Eco, 33.5%)
RP286 trbG conjugal transfer (B-Rsn, 24.7%)
RP103 virB4 virulence protein VIRB4 (B-Atu, 30.9%)
RP784 virB4 virulence protein VIRB4 (B-Atu, 20.3%)
RP287 virB8 virulence protein VIRB8 (B-Atu, 20.4%)
RP290 virB9 virulence protein VIRB9 (B-Atu, 24.8%)
RP291 virB10 virulence protein VIRB10 (B-Atu, 20.3%)
RP292 virB11 virulence protein VIRB11 (B-Atu, 29.6%)
RP293 virD4 virulence protein VIRD4 (B-Atu, 31.3%)

Drug and analogue sensitivity

RP170 acrD acriflavin resistance protein D (B-Eco, 31.3%)
RP475 ampG1 AMPG protein (B-Eco, 31.4%)
RP668 ampG2 AMPG prt (B-Eco, 26.3%)
RP781 ampG3 AMPG prt (B-Eco, 27.6%)
RP603 bcr1 bicyclomycin resistance (B-Eco, 21.7%)
RP698 bcr2 bicyclomycin resistance (B-Eco, 18.8%)
RP243 emrA multidrug resistance protein A (B-Eco, 26.9%)
RP157 emrB multidrug resistance protein B (B-Hin, 29.3%)
RP786 terC tellurite resistance protein (B-Eco, 35.3%)

Colicin-related functions

RP302 tolB colicin tolerance protein (B-Hin, 29.8%)
RP309 tolQ inner membrane protein (B-Eco, 39.7%)
RP310 tolR inner membrane protein (B-Pae, 40.1%)

Uncategorized

RP493 addA adducin alpha subunit (E-Hsa, 32.6%)
RP199 adx1 adrenodoxin precursor (E-Spo, 57.1%)
RP714 ank2 ankyrin (E-Hsa, 32.7%)
RP245 bolA BolA protein (B-Val, 34.2%)
RP181 claQ* thermostable carboxypeptidase (B-Pho, 29.1%)
RP297 cysQ sulphite synthesis pathway protein (B-Eco, 31.2
RP323 cyaY CyaY protein (B-Ech, 31.1%)
RP118 era GTP binding protein Era (B-Bsu, 33.6%)
RP063 hesB1 HesB protein (B-Ava, 37.0%)
RP484 hesB2 HesB protein (B-Pbo, 40.2%)
RP212 n2B N2B, ATPase protein (E-Hir, 27.2%)
RP485 nifU nitrogen fixation protein (B-Avi, 43.0%)
RP832 p34 P34 protein (B-Rri, 91.3%)
RP602 pat1 patatin B1 precursor protein (E-Stu, 22.9%)
RP317 pkcI protein kinase C inhibitor (B-Abr, 38.6%)
RP109 ptb phosphate buturyltransferase (B-Cab, 30.4%)
RP594 scoB* succinyl-CoA:3-ketoacid-CoA transferase subu (B-Mtu, 22.0%)
RP031 sco2 Sco yeast precursor protein (E-Sce, 32.6%)
RP587 sco2 Sco yeast precursor protein (E-Sce, 36.6%)
RP846 sfhB SfhB protein (B-Zmo, 40.6%)
RP430 smpB small protein (B-Syn, 46.7%)
RP058 soj SOJ protein (B-Bsu, 50.4%)
RP486 spl1 tRNA splicing protein (E-Cal, 58.9%)
RP487 spl1 tRNA splicing protein (E-Cal, 32.3%)
RP059 spo0J sporulation protein (B-Bsu, 40.4%)
RP239 suhB supressor protein (B-Eco, 22.8%)
RP733 surf1 SurF1 protein (E-Hsa, 23.9%)
RP710 tra3* transposase (B-Rme, 34.0%)

HYPOTHETICAL PROTEINS 1

Integral membrane proteins 19

NO SIMILARITY 2

Integral membrane proteins 21

Fig. 1). Comparison of the sequences from ten different *Rickettsia* species indicates that the disruption of the rRNA gene operon preceded the divergence of the typhus group and spotted fever group *Rickettsia* (S.G.E.A. *et al.*, unpublished observations). In addition, the genome contains a short sequence with similarity to a 13-nucleotide RNA molecule in *Bradyrhizobium japonicum* that may regulate transcription[25].

There are unusually few repeat sequences in this genome. We identified four different types of repeat sequence: all of these are located in intergenic regions. There is a sequence of 80 bp that is repeated seven times downstream of *rpmH* and *rnpA* in the *dnaA* region. A repetitive sequence of 325 bp is found at two intergenic regions that are more than 80 kb apart, downstream of the genes *ksgA* and *rnh*, respectively. A 440-bp-long repetitive sequence has been identified at two intergenic sites, 140 kb apart; one of these sites is downstream of *rrf* and the others downstream of *pdhA* and *pdhB*. Finally, two similar sequences of 730 bp are located immediately next to each other at 850 kb.

Paralogous families. We have identified 54 paralogous gene families comprising 147 gene products. Of these, 125 have an assigned function. Most paralogues encode proteins with transport functions, such as the ABC transporters, the proline/betaine transporters and the ATP/ADP transporters. Five paralogous genes located next to each other at 115 kb encode putative integral membrane proteins with unknown functions.

Table 1 Asterisks indicate putative pseudogenes. Abbreviations of species names are: Bacteria: *Acinetobacter calcoaceticus* (B-*Aca*), *Actonobacillus actinomycetemcomitans* (B-*Aac*), *Acyrthosiphon condii* (B-*Aco*), *Agrobacterium tumefaciens* (B-*Atu*), *Alcaligenes eutrophus* (B-*Aeu*), *Anabena sp.* PCC7120 (B-*Asp*), *Anabena variabilis* (B-*Ava*), *Anacystis nidulans* (B-*Ani*), *Azorhizobium caulinodans* (B-*Aca*), *Azospirillum brasiliense* (B-*Abr*), *Azotobacter vinelandii* (B-*Avi*), *Bacillus caldotenax* (B-*Bca*), *Bacillus stereothermophilus* (B-*Bst*), *Bacillus subtilis* (B-*Bsu*), *Bartonella bacilliformis* (B-*Bba*), *Bartonella henselae* (B-*Bhe*), *Bordetella pertussis* (B-*Bpe*), *Borrelia burgdorferi* (B-*Bbu*), *Bradyrhizobium japonicum* (B-*Bja*), *Brucella abortus* (B-*Bab*), *Brucella ovis* (B-*Bov*), *Caulobacter crescentus* (B-*Ccr*), *Chlamydia trachomatis* (B-*Ctr*), *Chloroflexus aurantiacus* (B-*Cau*), *Chromatum vinosum* (B-*Cvi*), citrus-greening-disease-associated bacterium (B-*Cgr*), *Clostridium acetobutylicum* (B-*Cac*), *Clostridium pasteurianum* (B-*Cpa*), *Clostridium thermosaccharolyticum* (B-*Cts*), *Coxiella burnetii* (B-*Cbu*), *Erwinia chrysanthemi* (B-*Ech*), *Escherichia coli* (B-*Eco*), *Haemophilus influenzae* (B-*Hin*), *Helicobacter pylori* (B-*Hpy*), *Klebsiella pneumoniae* (B-*Kpn*), *Legionella pneumophila* (B-*Lpn*), *Leucothrix mucor* (B-*Lmu*), *Liberobacter africanum* (B-*Laf*), *Methylobacterium extorquens* (B-*Mex*), *Micrococcus luteus* (*Mlu*), *Moraxella catarrhalis* (*Mca*), *Mycobacterium leprae* (*Mle*), *Mycobacterium smegmatis* (B-*Msm*), *Mycobacterium tuberculosis* (B-*Mtu*), *Mycoplasma capricolum* (B-*Mca*), *Mycoplasma genitalium* (B-*Mge*), *Mycoplasma pneumoniae* (B-*Mpn*), *Paracoccus denitrificans* (B-*Pde*), *Pasteurella haemolytica* (B-*Pha*), *Plectonema boryanum* (B-*Pbo*), *Proteus mirabilis* (B-*Pmi*), *Proteus vulgaris* (B*Pvu*), *Pseudomonas aeruginosa* (B-*Pae*), *Pseudomonas fluorescens* (B-*Pfl*), *Pseudomonas putida* (B-*Ppu*), *Pseudomonas syringae* (B-*Psy*), *Rhizobium meliloti* (B-*Rme*), *Rhizobium sp.* NGR234 (B-*Rsp*), *Rhodobacter capsulatus* (B-*Rca*), *Rhodobacter sphaeroides* (B-*Rsp*), *Rhodobacter sulfidophilus* (B-*Rsu*), *Rhodopseudomonas blastica* (B-*Rbl*), *Rhodospirillum rubrum* (B-*Rru*), *Rickettsia japonicum* (B-*Rja*), *Rickettsia rickettsii* (B-*Rri*), *Rickettsia typhi* (B-*Rty*), *Salmonella typhi* (B-*Sti*), *Salmonella typhimurium* (B-*Sty*), *Shigella flexneri* (B-*Sfl*), *Spiroplasma citri* (B-*Sci*), *Staphylococcus aureus* (B-*Sau*), *Staphyloccus carnosus* (B-*Sca*), *Streptococcus pneumoniae* (B-*Spn*), *Streptomyces clavuligerus* (B-*Scl*), *Streptomyces coelicor* (B-*Sco*), *Synechocystis* PCC 6803 (B-*Syn*), *Thermus aquaticus* (B-*Taq*), *Thermus thermophilus* (B-*Tth*), *Thiobacillus cuprinus* (B-*Tcu*), *Treponema hyodysenteriae* (B-*Thy*), *Vibrio alginolytics* (B-*Val*), *Vibrio cholera* (B-*Vch*), *Vibrio parahaemolyticus* (B-*Vpa*), *Vibrio proteolyticus* (B-*Vpr*), *Wolbachia* sp. (B-*Wsp*), *Yersinia entercolitica* (B-*Yen*), *Zooglea ramigera* (B-*Zra*), *Zymomonas mobilis* (B-*Zmo*). Archaea: *Methanococcus jannaschii* (A-*Mja*), *Sulfolobus acidocaldarius* (A-*Sac*). Eukaryotes: *Apis mellifera* (E-*Ame*), *Arabidopsis thaliana* (E-*Ath*), *Atratylodes japonica* (E-*Aja*), *Bos taurus* (E-*Bta*), *Candida albicans* (E-*Cal*), *Caenorhabidits elegans* (E-*Cel*), *Dictyostelium discoideum* (E-*Ddi*), *Flaveria trinervia* (E-*Ftr*), *Giardia theta* (E-*Gth*), *Glycine max* (E-*Gma*), *Haematobia irritans* (E-*Hir*), *Homo sapiens* (E-*Hsa*), *Marchantia polymorpha* (E-*Mpa*), *Mus musculum* (E-*Mmu*), *Prototheca wickerhamii* (E-*Pwi*), *Petunia hybrida* (E-*Phy*), *Pisum sativum* (E-*Psa*), *Porphyra purpurea* (E-*Ppu*), *Odontella sinensis* (E-*Osi*), *Reclinomonas americana* (E-*Ram*), *Rattus novergicus* (E-*Rno*), *Rhizopus oryzae* (E-*Ror*), *Saccharomyces cerevisiae* (E-*Sce*), *Schizosaccharomyces pombe* (E-*Spo*), *Solanum tuberosum* (E-*Stu*), *Spinacia oleracea* (E-*Sol*).

Biosynthetic pathways

A striking feature of the *R. prowazekii* genome is the small proportion of biosynthetic genes compared with free-living proteobacterial relatives (such as *Haemophilus influenzae*, *Helicobacter pylori* and *E. coli*)[15,19,20]. This scarcity of biosynthetic functions is also seen in diverse endocellular and epicellular parasites[16–18,23]. This scarcity of biosynthetic functions is also seen in diverse endocellular and epicellular parasites[16–18,23].

Amino-acid metabolism. As many as 43 and 69 genes required for amino-acid biosynthesis are found in *Helicobacter pylori* and *Haemophilus influenzae*, respectively. In contrast, *Mycoplasma genitalium* and *Borrelia burgdorferi* contain only *glyA*, which encodes serine hydroxymethyltransferase. This gene is also found in *R. prowazekii* (Table 1). Serine hydroxymethyltransferase catalyses the conversion of serine and tetrahydrofolate into glycine and methylenetetrahydrofolate, respectively. A role in tetrahydrofolate metabolism may account for the ubiquity of *glyA* in bacteria.

Seven genes normally associated with lysine biosynthesis (*lysC*, *asd*, *dapA*, *dapB*, *dapD*, *dapE* and *dapF*) are also present in *R. prowazekii*. The biosynthetic pathways leading to lysine, methionine and threonine share the first two of these (*lysC* and *asd*). However, none of the downstream genes for threonine biosynthesis are found in *R. prowazekii*. Likewise, the lysine pathway is incomplete, and *lysA*, which encodes the enzyme that converts meso-diaminopimelate to lysine, is missing. The likely role of the upstream genes of this pathway in *R. prowazekii* is the biosynthesis of diaminopimelate, an essential envelope component. We have therefore classified these genes as 'cell-envelope' genes (Table 1).

We have identified other genes that are superficially involved in the metabolism of amino acids, but which apparently function in deamination pathways that divert amino acids into the tricarboxylic acid (TCA) cycle. For example, there is *aatA*, encoding aspartate aminotransferase, which catalyses the degradation of aspartate to oxaloacetate and glutamate. *tdcB* encodes threonine deaminase, which converts threonine into α-ketobutyrate. Another gene (*ilvE*) encodes branched-chain-amino-acid aminotransferase, which converts leucine, isoleucine or valine into glutamate. *pccA* and *pccB* encode propionyl-CoA carboxylase, which converts propionyl-CoA, an intermediate in the breakdown of methionine, valine and isoleucine, into succinyl-CoA. The *pccA* and *pccB* gene products show greatest similarity to the eukaryotic proteins that are located in the mitochondrial matrix.

Nucleotide biosynthesis. No genes required for the *de novo* syntheses of nucleosides have been found in the *R. prowazekii* genome. However, four genes required for the conversion of nucleoside monophosphates into nucleoside diphosphates (*adk*, *gmk*, *cmk* and *pyrH*) are present. There are also two genes encoding ribonucleotide reductase, which converts ribonucleoside diphosphates into deoxyribonucleoside diphosphates. Nucleoside diphosphate kinase (encoded by *ndk*), which converts NDPs and dNDPs to NTPs and dNTPs, is also present in *R. prowazekii*. Finally, there is a complete set of genes for the conversion of dCTP and dUTP into TTP, including *thyA*, which codes for thymidylate synthase. Thus, the *R. prowazekii* genome encodes all of the enzymes required for the interconversion of nucleoside monophosphates into all of the other required nucleotides. The nucleoside monophosphates are probably imported from the eukaryotic host.

Energy metabolism

Early in its infectious cycle, *R. prowazekii* uses the ATP of the host with the help of membrane-bound ATP/ADP translocases. However, *R. prowazekii* is also capable of generating ATP, which may compensate for the gradual depletion of cytosolic ATP later in the infection. *R. prowazekii*'s repertoire of genes involved in ATP production and transport include determinants for the TCA cycle, the respiratory-chain complexes, the ATP-synthase complexes and the ATP/ADP translocases (Table 1). Genes to support anaerobic glycolysis are absent.

Pyruvate dehydrogenase. Pyruvate is imported into mitochondria directly from the cytoplasm and converted into acetyl-CoA by pyruvate dehydrogenase. The genes encoding three components (E1–E3) of the pyruvate dehydrogenase complex are found in *R. prowazekii*, indicating that it too uses cytosolic pyruvate. Pyruvate dehydrogenase (E1) consists of two subunits (α and β) in *R. prowazekii*, mitochondria and Gram-positive bacteria; the corresponding genes are clustered in the genome. In contrast, proteobacteria such as *E. coli*, *Haemophilus influenzae* and *Helicobacter pylori* have a single subunit for the E1 component and these have little similarity to the α and β subunits of the E1 component in *R. prowazekii* and mitochondria (data not shown).

Two paralagous genes code for the dihydrolipoamide dehydrogenase (E3) in *R. prowazekii*. One of these most resembles mitochondrial homologues, whereas the other is most similar to bacterial homologues (data not shown). The presence of several paralogous gene families for pyruvate dehydrogenases complicates attempts to reconstruct a genome phylogeny based on these genes.

ATP production. Genes encoding all enzymes in the TCA cycle are found in *R. prowazekii*. Proton translocation is mediated by NADH dehydrogenase (complex I), cytochrome reductase (complex III) and cytochrome oxidase (complex IV). Several clusters of genes code for components of the NADH dehydrogenase complex. Seven of these genes (*nuoJKLM* and *nuoGHO*) are located near to each other, but the order of genes is inverted relative to the order of this cluster in *E. coli*. An additional set of five genes is grouped in the order *nuoABCDE*, but the single genes *nuoF* and *nuoN* are distant from both of these clusters. Several proteins in the cytochrome bc_1 reductase complex, such as ubiquinol–cytochrome *c* reductase (encoded by *petA*), cytochrome *b* (encoded by *cytb*) and cyto- chrome c_1 (encoded by *fbhC*), are present, as are several subu[nits of] the cytochrome oxidase complex.

The ATP-synthesizing complex is composed of the ATP syn[thase] F_1 component (comprising five polypeptides, α, β, γ, ϵ and δ) [and] the F_o component, a hydrophobic segment that spans the mitochondrial membrane. The genes encoding these compo[nents] are normally clustered in one of the most highly conserved o[peron] structures in microbial genomes. In *R. prowazekii*, howeve[r, the] ATP-synthase genes encoding the α, β, γ, δ and ϵ subunits of [the] ATP-synthase genes encoding the α, β, γ, δ and ϵ subunits of the complex (*atpH*, *atpA*, *atpG*, *atpD* and *atpC*) are clustered [in the] common order, but *atpB*, *atpE* and *atpF*, encoding the A, B [and] chains of the F_o complex, are split from this cluster.

Replication, repair and recombination

R. prowazekii has a smaller set of genes involved in DNA replic[ation] than do free-living bacteria such as *E. coli*, *Haemophilus influ[enzae]* and *Heliocobacter pyrlori*. Four genes have been identified tha[t code] for the core structure of DNA polymerase III, which includes [the α] (*dnaE*), ϵ (*dnaQ*), β (*dnaN*), γ and θ (*dnaX*) subunits[. Other] subunits present in the *E. coli* DNA polymerase III are mi[ssing] from *R. prowazekii*, as well as from *M. genitalium* and *B. burgdo[rferi]*.

Genes encoding DNA-repair mechanisms are similar in the genomes of the parasites *R. prowazekii*, *M. genitalium* [and] *B. burgdorferi*. Thus, genes involved in the repair of ultravi[olet-] induced DNA damage (*uvrABCD*) have been identified in all [three] genomes. In *R. prowazekii*, DNA-excision repair probably occu[rs by] a pathway involving endonuclease III, polI and DNA ligase, [as in] *B. burgdorferi*.

The *R. prowazekii* genome has a limited capacity for mism[atch] repair. The DNA-mismatch-repair enzymes encoded by *mutL* and [*mutS*] are present, but *mutH* and *mutY* are not. There is a complete lack o[f these] genes in *M. genitalium*, but *mutL* and *mutHLY* have been iden[tified] in *B. burgdorferi* and *Chlamydia trachomatis*. The transcrip[tion-] repair coupling factor (encoded by *mfD*) is found in *R. prowaz[ekii]*, *B. burgdorferi* and *C. trachomatis* but not in *M. genitalium*.

The *R. prowazekii* genome contains several genes involve[d in] homologous recombination, such as *recA*, *recF*, *recJ*, *recN* and [...]. A similar set of genes has been found in *A. aeolicus*[21]. The rec[...]

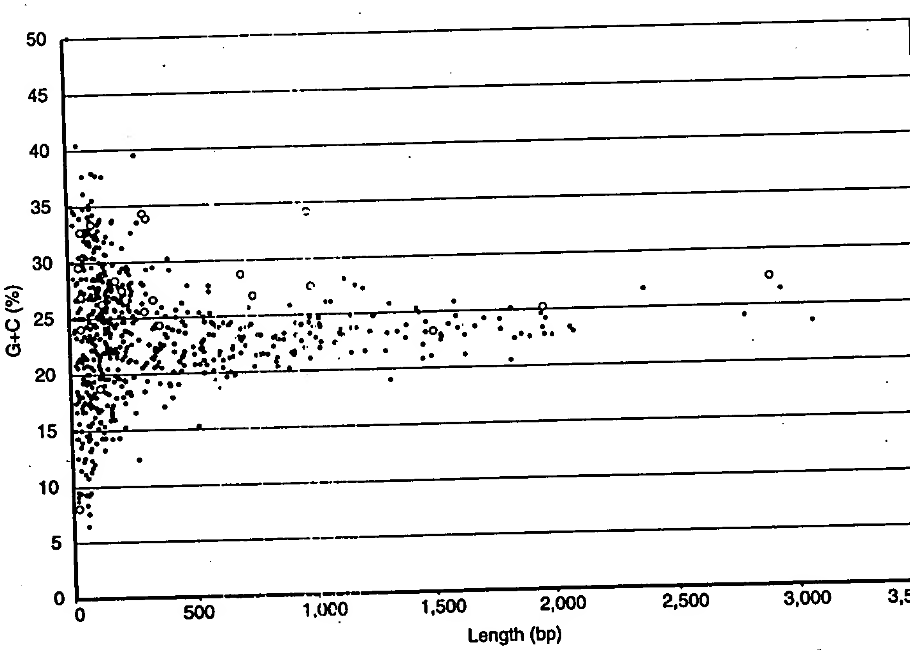


Figure 3 G+C content in intergenic regions longer than 20 bp in the *R. prowazekii* genome. The empty circles correspond to spacer sequences located at 886 to 916 kb, a region with an unusually large fraction of non-coding DN[A and] pseudogenes.

scattered in the other small genomes of parasites. The *RecBCD* complex is missing in *R. prowazekii*, *M. genitalium* and *Helicobacter pylori* but it has been identified in *B. burgdorferi*.

Transcription and translation

R. prowazekii has three subunits (α, β and β′) of the core RNA polymerase, as well as σ^{70} and one alternative σ factor, σ^{32}, which controls transcription of the genes encoding heat-shock proteins in *E. coli*. Genes involved in transcription elongation and termination, *nusA*, *nusB*, *nusG*, *greA* and *rho*, are also present. The gene encoding *nusA* is absent in most other small genomes, such as those of *B. burgdorferi*, *Helicobacter pylori*, *M. genitalium* and *C. trachomatis*, although genes for heat-shock proteins are present.

An unusually large number of genes involved in RNA degradation are found in *R. prowazekii*. Of these, only four appear to be common to the bacterial genomes analysed so far (those encoding polyribonucleotide nucleotidyltransferase and ribonucleases HII, III and P). Four more ribonucleases (D, E, HI and PH) are present in *R. prowazekii*, but in none of the other small parasites.

Of the 33 identified tRNA genes, which code for 32 different tRNA isoacceptor species, two code for $tRNA^{Phe}$. There are two tRNA species for most of the amino acids that are encoded by four-codon boxes; the exceptions are the four-codon boxes for proline and valine, for which we have identified only one isoacceptor-tRNA species, with U in the first anticodon position. *selC*, which codes for $tRNA^{Sec}$, and *selABD* are missing. *R. prowazekii* has a set of genes coding for tRNA modifications (*tgt*, *queA*, *trmD*, *truA*, *truB* and *miaA*) which resembles that of *Helicobacter pylori*, *C. trachomatis* and *B. burgdorferi*; *M. genitalium* has only *trmD* and *truA*.

In *R. prowazekii*, 21 genes encode 18 of the 20 aminoacyl-tRNA synthetases normally required for protein synthesis. There are two genes (*gltX*) encoding glutamyl-tRNA synthetase. As seen in several bacterial genomes[25], the gene coding for glutaminyl-tRNA synthetase, *glnS*, is missing. Three genes encoding subunits of the glutamyl-tRNA amidotransferase are present, indicating that a glutamyl-tRNA charged with glutamic acid may be transamidated to generate Gln-tRNA. The gene coding for asparaginyl-tRNA synthetase, *asnS*, is also missing from the *R. prowazekii* genome as well as from *Helicobacter pylori*, *C. trachomatis* and *A. aeolicus*[26]. A transamidation process to form $Asn\text{-}tRNA^{Asn}$ from $Asp\text{-}tRNA^{Asn}$ has been proposed for the archaeon *Haloferax volcanii*[27] and this reaction may also occur in *R. prowazekii*. The valyl-tRNA synthetase is 38.3% identical to its homologue in *Methanococcus jannaschii*, but only 27.6% identical to its most similar homologue in bacteria, which is found in *Bacillus stearothermophilus*, possibly indicating a horizontal transfer event. The lysyl-tRNA synthetase (encoded by *lysS*) in *R. prowazekii* is a class I enzyme with no resemblance to the conventional class II lysyl-tRNA synthetases. Class I type of lysyl-tRNA synthetases have been observed previously in only *B. burgdorferi*, *Pyrococcus woesii*, *Methanococcus jannaschii* and a few other methanogens[26].

Regulatory systems

As in other genomes of small parasites, *R. prowazekii* has a reduced set of regulatory genes. There are a few members of two-component regulatory systems, such as the proteins encoded by *barA*, *envZ*, *ntrY*, *ntrX*, *ompR* and *phoR*. *spoT*, which is involved in the stringent response, has been identified in *B. burgdorferi*, *Helicobacter pylori* and *M. genitalium*. Only remnants of genes coding for amino-terminal fragments of proteins similar to those encoded by *spoT* and *relA* are identifiable in *R. prowazekii*. No fragments of *spoT* encoding the carboxy-terminal segments of the protein have been identified in the genome.

Cell division and protein secretion

Proteins involved in detoxification, such as superoxide dismutase, and those involved in thiophen and furan oxidation are present in *R. prowazekii*. Two genes encoding haemolysins have also been identified, and an *R. typhi* homologue of *tlyC* exhibits haemolytic activities when expressed in *E. coli* (S. Radulovic, J. M. Troyer, B. Noden, S.G.E.A. and A. Azad, unpublished observations).

The data indicate that the basic mechanisms of cell division and secretion in *R. prowazekii* are similar to those in free-living proteobacteria. There is a common set of bacterial chaperones (encoded by *dnaK*, *dnaJ*, *hslU*, *hslV*, *groEL*, *groEL*, *groES* and *htpG*) and genes involved in the *secA*-dependent secretory system (*secABDEFGY*, *ffH* and *ftsY*). *R. prowazekii* has a significantly larger set of genes involved in peptide secretion than does *M. genitalium*.

Membrane-protein analysis

Many studies of *R. prowazekii* have focused on outer-surface membrane proteins because of their potential importance in bacterial detection and vaccination. The superficial lipopolysaccharide (LPS) molecule is important in the pathogenesis of *R. prowazekii*. LPS consists of a polysaccharide that is covalently linked to lipid A, the biosynthesis of which is catalysed by products of *lpxABCD*, all of which are present in the *R. prowazekii* genome. These genes are clustered in *E. coli*, but *lpxA* and *lpxD* are separate from *lpxB* and *lpxC* in *R. prowazekii*. Three genes involved in the biosynthesis of the 3-deoxy-D-manno-octulosonic acid (KDO) residues reside in the *R. prowazekii* genome (*kdsA*, *kdsB* and *kdtA*). Only one gene (*rfaJ*) with a putative function in outer-core biosynthesis has been identified.

We have identified a set of genes involved in the biosynthesis of murein and diaminopimelate and a set involved in the biosynthesis of fatty acids. These includes: *fabD*, which is involved in the last step of the initiation phase of fatty-acid biosynthesis; four genes involved in the elongation cycle of fatty-acid biosynthesis (*fabFGHI*); and three genes involved in the first three steps of the synthesis of polar head groups (*cdsA*, *pssA* and *pgsA*). Finally, post-translational processing and addition of lipids to an N-terminal cysteine require the gene products prolipoprotein diacylglycerol transferase (*lgt*), prolipoprotein signal peptidase (*lspA*) and apolipoprotein:phospholipid *N*-acyl transferase (*lnt*). These are found in the genome with several genes involved in the degradation of fatty acids, such as *fadA* which encodes the 3-ketoacyl-CoA thiolase.

Virulence

The *R. prowazekii* genome contains several homologues of the *VirB* gene operon found in *Agrobacterium tumefaciens*. This gene family encodes proteins that direct the export of the T-DNA–protein complex across the bacterial envelope to the plant nuclei[28]. *R. prowazekii* has two homologues of *VirB4* and one homologue each of *VirB8*, *VirB9*, *VirB10*, *VirB11* and *VirD4*. The latter five genes are clustered with the gene *trbG*, which is involved in conjugation in *Agrobactrium tumefaciens*. Homologues of the single-stranded DNA-binding proteins VirD2 and VirE2 are missing. In *Agrobacterium tumefaciens*, these proteins are bound to the transferred T-DNA, indicating different functions for the homologues of the *VirB* genes in *R. prowazekii*. Indeed, VirB proteins are homologous to components of the *E. coli* transport system for plasmids, as well as to components of the Pt1 transport machinery in *Bordetella pertussis*, which exports pertussis toxin[28]. A set of genes coding for VirB4 and several other VirB proteins has been identified in the *cag* pathogenicity island of *Helicobacter pylori*. In this species, the VirB proteins facilitate export of a factor that induces interleukin-8 secretion in gastric epithelial cells[28]. Thus, *R. prowazekii* may encode components of a transport system for both conjugal DNA transfer and protein export.

The virulence of *Staphylococcus aureus* has been correlated with the production of capsular polysaccharides in phagocytic assays and mouse lethality assays[29,30]. A cluster of ten capsule genes (*capA–M*) is involved in capsule biosynthesis in *S. aureus* strain M[31]. We have identified three *R. prowazekii* genes with sequence similarities to *S. aureus cap* genes. Two of these (*capD* and *capM*) are separated by ten

genes, most of which are unknown genes or genes involved in the biosynthesis of LPS or techoic acid. Thus, *R. prowazekii* may produce components of a microcapsular layer that is involved in virulence.

Reductive evolution

Genome sequences of organisms enjoying an endosymbiotic lifestyle are at risk. The activities of homologous nuclear genes may render genes of the endosymbiont expendable and as a consequence they become vulnerable to obliteration by mutation. Good candidates for such purged genes in *Rickettsia* and mitochondria are genes required for amino-acid biosynthesis, nucleoside biosynthesis and anaerobic glycolyis. These and other genes would have been deleted when an ancestral genome first lived in a nucleated cell. Once genes essential to a free-living mode are lost, the endosymbiont becomes an obligate resident of its host.

Likewise, small, bottle-necked populations of bacteria infecting a eukaryotic cell will tend to accumulate deleterious mutations because selection cannot remove them from such clonal populations[13]. The accumulation of such harmful but non-lethal mutations is referred to as 'Muller's ratchet'[32] or 'near-neutral evolution'[33,34]. The consequence of accumulation of these mutations will be the inactivation and eventual deletion of non-essential genes.

The first mutation that inactivates an expendable gene is likely to initiate a sequence of events in which subsequent mutations freely transform it, by degrees, from a pseudogene, to unrecognizable sequence, to small fragments, to extinction. In this sequence, mutations are released from amino-acid-coding constraints. Thus nucleotide substitutions will reflect the mutation bias of the genome. This bias can be estimated roughly by frequencies of third-position bases in the codons. For *R. prowazekii*, the bias of the third-position bases is 18% G+C rather than the 29% G+C average for the genome. So, as sequences age in *R. prowazekii*, their composition should gradually approach the low G+C content of third codon positions. Nearly one-quarter of the *R. prowazekii* genome is composed of non-coding sequences, with a G+C content lower than that of coding sequences (25% G+C compared to 30%; $P < 0.001$). Thus, much of the non-coding sequence may be remnants of coding sequences that are in the process of being eliminated from the genome.

The gene encoding *S*-adenosylmethionine synthetase (*metK*), which catalyses the biosynthesis of *S*-adenosylmethionine (SAM), illustrates the initiation of this process. The *metK* sequence in the strain of *R. prowazekii* studied here has a termination codon within a region of the gene that is otherwise highly conserved among bacterial species[35]. However, a closely related strain does not the termination codon. Many other defects, such as termin codons, insertions, and a preponderance of small deletions also been observed in the *metK* genes in several members of spotted fever group *Rickettsia* (J.O.A. and S.G.E.A., unpubli observations). This random distribution of lethal mutations am some *metK* alleles from different *Rickettsia* species indicates th gene may have just entered the extinction process. This distribu and the identification of 11 more pseudogenes for carboxypepti (*ypwA*), penicillin-binding protein (*pbpC*), succinyl CoA-tran ase (*scoB*), transposase (*tra3*), resolvase (*pin*), conjugative tra protein (*taxB*), a hypothetical protein (*yfc1*) and four diff fragmented open reading frames for (p)ppGpp 3′-pyrophosph drolase, indicates that the *R. prowazekii* genome continu eliminate genes.

Genome sequences can be purged by a more abrupt mechan This consists of intrachromasomal recombination at dupli sequences, which can result in the deletion of intervening seque the loss of a sequence duplication and the rearrangement of flan

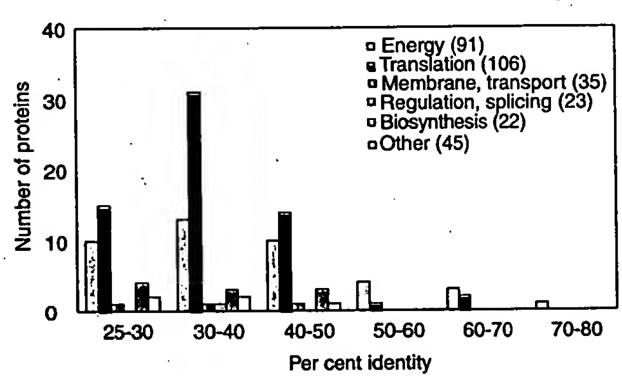


Figure 4 Histogram representation of the similarity of predicted *R. prowazekii* proteins to yeast proteins targeted to the mitochondria. Only protein pairs with per cent identity values greater than 25% are shown. Numbers in parentheses represent the total number of yeast mitochondrial proteins within each category. The yeast mitochondrial protein sequences have been taken from http://www.proteome.com.

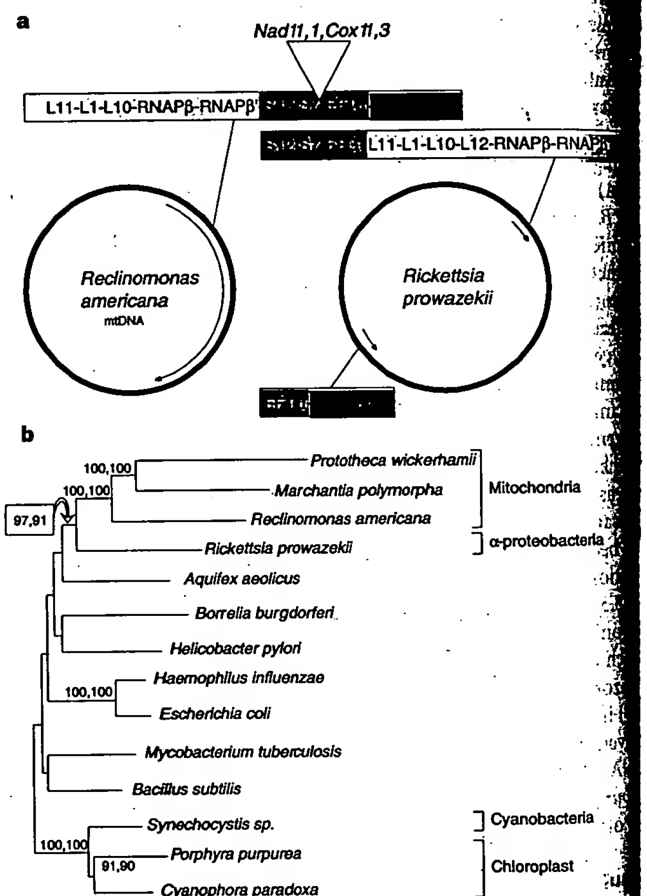


Figure 5 The organization and phylogenetic relationships of gene enco ribosomal protein from *R. prowazekii* and the mitochondrial genom *Reclinomonas americana*. **a**, The organization of ribosomal-protein genes S10, *spc* and α-operons are organized similarly in these two genomes, excep several ribosomal-protein genes[38] have been deleted from the mitochon genome of *Reclinomonas americana*. **b**, The phylogenetic relationship mitochondria and bacteria were derived from the combined amino sequences of ribosomal proteins S2, S3, S7, S10, S11, S12, S13, S14, S19, and L16. Neighbour-joining and maximum-parsimony methods gave ide topologies. Branch lengths are proportional to those reconstructed by usin neighbour-joining method. Values at nodes are bootstrap values indicatin degree of support for individual clusters under each method (neighbour-jo maximum parsimony). Only bootstrap values >90% are shown.

sequences. Such a mechanism will account for the presence in *R. prowazekii* of one, unlinked copy of *rrs* and *rrl*, both of which are surrounded by new flanking sequences[36]. Likewise, *R. prowazekii* has one *tuf* gene and one *fus* gene in atypical clusters that seem to have been created by intrachromosomal recombination between the two *tuf* genes that are normally found in Gram-negative bacteria[37]. Indeed, rearranged gene operon structures encoding ribosomal proteins are characteristic of all members of the genus *Rickettsia* (H. Amiri, C.A. and S.G.E.A., unpublished observations).

Conserved operons that are found in free-living bacteria are often dispersed throughout the *Rickettsia* genome (see above). The *R. prowazekii* genome contains an unusually small fraction of repeat sequences (<10% of that observed in free-living bacteria). We suggest that the repeat sequences found in the ancestor to the *Rickettsia* have been 'consumed' by the intrachromosomal-recombination mechanism that generated some of the deletions and rearrangements seen in *R. prowazekii*. Such intrachromosomal recombinants arise at a substantial rate in bacteria growing in culture, but here they are eliminated from the populations by selection. That such remnants of intrachromosomal recombination are retained in *R. prowazekii* indicates that purifying selection has been attenuated in this organism.

Mitochondrial affinities

The reduction in genome size in mitochondria and *Rickettsia* is likely to have occurred independently in the two lineages. Most of the genes supporting mitochondrial activities are nuclear. Many of the 300 proteins encoded in the nucleus of the yeast *Saccharomyces cerevisiae* but destined for service within the mitochondrion are close homologues of their counterparts in *R. prowazekii*. Nearly one-quarter of these proteins are required for bioenergetic processes and another one-third of them are required for the expression of the genes encoded in the mitochondrial genome. In total, more than 150 nucleus-encoded mitochondrial proteins share significant sequence homology with *R. prowazekii* proteins (Fig. 4).

Another group of 58 nucleus-encoded mitochondrial proteins represents components of the mitochondrial transport machinery and regulatory system (Fig. 4). These include proteins found in the mitochondrial outer membrane and others involved in splicing reactions. Such proteins have probably been secondarily recruited to mitochondria from genomes not necessarily related to that of the α-proteobacterial ancestor.

The mitochondrial genome of the early diverging, freshwater protozoan *Reclinomonas americana* is more like that of a bacterium than any other mitochondrial genome sequenced so far[38]. This genome contains 67 protein-coding genes, most of which provide components of genetic processes and the bioenergetic system[38]. Several gene clusters in this mitochondrial genome are reminiscent of those in bacteria (Figs 5a, 6a). Most similarities represent retained, ancestral traits present in the common ancestor of bacteria and mitochondria. For example, the genes *rplKAJL* and *rpoBC* are identically organized in *R. prowazekii* and the mitochondrial genome of *Reclinomonas americana*. Likewise, the genes encoding the S10, *spc* and the α-ribosomal protein operons are organized similarly in the two genomes. The immediate proximity of these two clusters in the *Reclinomonas americana* mitochondrial DNA is reminiscent of the arrangement in free-living bacteria, whereas the physical separation of the two clusters in the *R. prowazekii* genome is atypical. A further rearrangement event is indicated by the fact that the *rpsLrpsGfus* cluster is located upstream of the *rplKAJLrpoBC* cluster in *R. prowazekii*, rather than downstream as it is in the *Reclinomonas americana* mtDNA. Phylogenetic reconstructions based on ribosomal proteins within each of these two clusters indicate that there is a close evolutionary relationship between *R. prowazekii* and mitochondria (Fig. 5b).

Mitochondria and *R. prowazekii* have a similar repertoire of proteins involved in ATP production and transport, including genes encoding components of the TCA cycle, the respiratory-chain complexes, the ATP-synthase complexes and the ATP/ADP translocases. There are some similarities in the gene orders of some functional clusters (Fig. 6a). There are also some rearrangements of clusters that are specific to *Rickettsia*. One example is the inversion of segments corresponding to *nuoJKLM* and *nuoGHI*. Another is the scattered displacement of genes involved in the biogenesis of cytochrome *c*. Nevertheless, phylogenetic reconstructions based on components of the NADH dehydrogenase complexes indicate that there is a close evolutionary relationship between *R. prowazekii* and mitochondria (Fig. 6b).

We have identified as many as five genes coding for ATP/ADP transporters, all of which are expressed (R.M.P. *et al.*, unpublished observations). The *Rickettsia* ATP/ADP translocases are monomers with 12 transmembrane regions each, whereas the mitochondrial translocates are dimers with six transmembrane regions per dimer. We found no relationship between the primary structures of the mitochondrial and *Rickettsia* ATP/ADP translocases, indicating that these transport systems may have originated independently.

The study of the *R. prowazekii* genome sequence supports the idea that aerobic respiration in eukaryotes originated from an ancestor of the *Rickettsia*, as indicated previously by phylogenetic reconstructions based on the rRNA gene sequences[7,9]. Phylogenetic analyses of the *petB* and *coxA* genes indicate that the respiration systems of *Rickettsia* and mitochondria diverged ~1,500–2,000 million years ago[10], shortly after the amount of oxygen in the atmosphere began

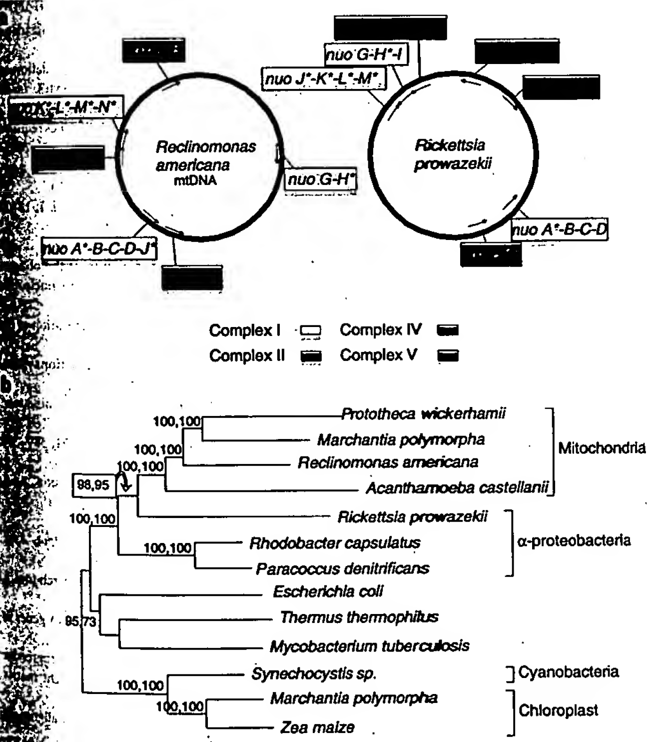


Figure 6 The organization and phylogenetic relationships of genes involved in ATP synthesis from *R. prowazekii* and the mitochondrial genome of *Reclinomonas americana*. **a**, The organization of bioenergetic genes. **b**, The phylogenetic relationships of mitochondria and bacteria were derived from the combined amino-acid sequences of NADH dehydrogenase I chains A, J, K, L, M and N, which are encoded by the genes *nuoA, J, K, L, M, N*. These genes are highlighted by asterisks in **a**. Neighbour-joining and maximum-parsimony methods gave identical topologies. Branch lengths are proportional to those constructed using the neighbour-joining method. Values at nodes are bootstrap values indicating the degree of support for individual clusters under each method (neighbour-joining, maximum parsimony). Only bootstrap values >90% are shown.

to increase. The finding that the ATP/ADP translocases in *R. prowazekii* and mitochondria are of different evolutionary origin is problematic (R.M.P. *et al.*, unpublished observations). Free-living bacteria do not seem to have homologues of ATP/ADP translocases, which are found only in organelles and in two obligate intracellular parasites, *Rickettsia* and *Chlamydia*. Thus it is not known whether the original endosymbiont was capable of efficient exchange of adenosine nucleotides with its host cell. More detailed comparative analysis of the genomes of α-proteobacteria may refine our understanding of the origins of mitochondria. □

Methods

Genome sequencing. We prepared genomic DNA from the Madrid E strain of *R. prowazekii*, which was originally isolated in Madrid from a patient who died in 1941 with epidemic typhus. We propagated *R. prowazekii* in the yolk sac of embryonated hen eggs and purified DNA according to standard procedures[39]. We sequenced the *R. prowazekii* genome by a whole-genome shotgun approach in combination with shotgun sequencing of a selected set of clones from a cosmid library (A.Z. *et al.*, unpublished observations). Genomic and cosmid DNA was sheared by nebulization to an average size of ~2 kb. The random fragments were cloned into a modified M13 vector using the double adaptor method[40]. We collected 19,078 sequence reads during the random sequencing phase using Applied Biosystems 377 DNA sequencers (Perkin-Elmer).

The sequences were assembled and the consensus sequence was edited using the STADEN program[41]. We verified the structure of the assembled sequence by end-sequencing of 3-kb-insert λ Zap II clones[36], 10-kb λ clones and 30-kb cosmid clones. More than 97% of the genome was covered by clones from the three different libraries (A.Z. *et al.*, unpublished observations). Gaps between contigs were closed by direct sequencing of clones from the three libraries or of polymerase chain reaction (PCR) products. The final four gaps were closed by direct sequencing of PCR products generated with the Long Range PCR system (Gene Amp). Regions of ambiguity were identified by visual inspection of the assembly and resequenced. The final assembly contains ~20,000 sequences. The genome sequence has eightfold coverage on average and no single region has less than twofold coverage. We estimate the overall error frequency to be $<1 \times 10^{-5}$.

Informatics. Sequence analysis and annotation was managed by CapDB (T.S.-P. *et al.*, unpublished observations). We identified open reading frames of more than 50 codons as genes on the basis of their characteristic patterns in nucleotide-frequency statistics[14] using BioWish[42]. The identified genes were analysed using the program BLASTX[43] to search for sequence similarities in EMBL, TREMBL, SwissProt and in-house databases. We identified tRNA genes with the program tRNA scan-SE[44]. Remaining frameshifts were considered to be authentic and annotated as pseudogenes. Families of paralogues were constructed using BLAST to search for sequence similarities within the *R. prowazekii* genome. Multiple alignments and phylogenetic trees for genes with significant sequence similarities to genes in the public databases were constructed automatically using CLUSTAL-W[45], Phylo_win[46] and GRS[47]. The final annotation was based on manual inspection of the phylogenetic placement of *R. prowazekii* in the resulting gene trees.

Received 21 July; accepted 24 September 1998.

1. Gross, L. How Charles Nicolle of the Pasteur Institute discovered that epidemic typhus is transmitted by lice: reminiscences from my years at the Pasteur Institute in Paris. *Proc. Natl Acad. Sci. USA* 93, 10539–10540 (1996).
2. Weisburg, W. G., Woese, C. R., Dobson, M. E. & Weiss, E. A common origin of Rickettsiae and certain plant pathogens. *Science* 230, 556–558 (1985).
3. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* 51, 221–227 (1987).
4. Weisburg, W. G. *et al.* Phylogenetic diversity of the rickettsias. *J. Bacteriol.* 171, 4202–4206 (1989).
5. Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. Mitochondrial origins. *Proc. Natl Acad. Sci. USA* 82, 4443–4447 (1985).
6. Gray, M. W., Cedergren, R., Abel, Y. & Sankoff, D. On the evolutionary origin of the plant mitochondrion and its genome. *Proc. Natl. Acad. Sci. USA* 86, 2267–2271 (1989).
7. Olsen, G. J., Woese, C. R. & Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* 176, 1–6 (1994).
8. Viale, A. & Arakaki, A. K. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett.* 341, 146–151 (1994).
9. Gray, M. W. & Spencer, D. F. in *Evolution of Microbial Life* (eds Roberts, D. M., Sharp, P. M., G. & Spencer, D. F.) 109–126 (Cambridge Univ. Press, Cambridge, 1996).
10. Sicheritz-Pontén, T., Kurland, C. G. & Andersson, S. G. E. A phylogenetic analysis of the cyto and cytochrome *c* oxidase I genes supports an origin of mitochondria from within the Ricke *Biochim. Biophys. Acta* 1365, 545–551 (1998).
11. Margulis, L. *Origin of Eukaryotic Cells* (Yale Univ. Press, New Haven, 1970).
12. Kurland, C. G. Evolution of mitochondrial genomes and the genetic code. *Bioessays* 14 (1992).
13. Andersson, S. G. E. & Kurland, C. G. Reductive evolution of resident genomes. *Trends Mi* 263–268 (1998).
14. Andersson, S. G. E. & Sharp, P. M. Codon usage and base composition in *Rickettsia prowaze* *Evol.* 42, 525–536 (1996).
15. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus* Rd. *Science* 269, 496–511 (1995).
16. Fraser, C. M. *et al.* The *Mycoplasma genitalium* genome reveals a minimal gene complem 270, 397–403 (1995).
17. Himmelreich, R. *et al.* Complete genome sequence analysis of the genome of the *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 3, 109–136 (1996).
18. Fraser, C. M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Na* 580–586 (1997).
19. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori* 388, 539–547 (1997).
20. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 14 (1997).
21. Deckert, G. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Na* 353–358 (1998).
22. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete sequence. *Nature* 393, 537–544 (1998).
23. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete 281, 375–388 (1998).
24. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol.* 660–665 (1996).
25. Ebeling, S., Kündig, C. & Hennecke, H. Discovery of a rhizobial RNA that is essential for root nodule development. *J. Bacteriol.* 173, 6373–6382 (1991).
26. Koonin, E. V. & Aravind, L. Genomics: re-evaluation of translation machinery evolution. *Curr.* 266–269 (1998).
27. Curnow, A. W., Ibba, M. & Soll, D. tRNA-dependent asparagine formation. *Nature* 382, (1996).
28. Christie, P. J. *Agrobacterium tumefaciens* T-complex transport apparatus: a paradigm for a new of multifunctional transporters in eubacteria. *J. Bacteriol.* 179, 3085–3094 (1997).
29. Melly, M. A., Duke, L. J., Liau, D.-F. & Hash, J. H. Biological properties of the encaps *Staphylococcus aureus*. *M. Infect. Immun.* 10, 389–397 (1974).
30. Peterson, P. K., Wilkinson, B. J., Kim, Y., Schmeling, D. & Quie, P. G. Influence of encapsulation staphylococcal opsonization and phagocytosis by human polymorphonuclear leukocytes. *Immun.* 19, 943–949 (1978).
31. Lin, W. S., Cunnen, T. & Lee, C. Y. Sequence analysis and molecular characterization of genes re for the biosynthesis of type I capsular polysaccharide in *Staphylococcus aureus*. *J. Bacteriol.* 176, 7016 (1994).
32. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* 78, 157–159 (1977).
33. Ohta, T. & Kimura, M. On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* 1, (1971).
34. Ohta, T. Evolutionary rate of cistrons and DNA divergence. *J. Mol. Evol.* 1, 150–157 (1972).
35. Andersson, J. O. & Andersson, S. G. E. Genomic rearrangements during evolution of the ob intracellular parasite *Rickettsia prowazekii* as inferred from an analysis of 52 015 bp nu sequence. *Microbiology* 143, 2783–2795 (1997).
36. Andersson, S. G. E., Zomorodipour, A., Winkler, H. H. & Kurland, C. G. Unusual organization of rRNA genes in *Rickettsia prowazekii*. *J. Bacteriol.* 177, 4171–4175 (1995).
37. Andersson, S. G. E. & Kurland, C. G. Genomic evolution drives the evolution of the translation *Biochem. Cell Biol.* 73, 775–787 (1995).
38. Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial genome in mini *Nature* 387, 493–497 (1997).
39. Winkler, H. H. Rickettsia permeability: an ATP/ADP transport system. *J. Biol. Chem.* 251, 38 (1976).
40. Andersson, B. *et al.* A 'double adaptor' method for improved shotgun library construction *Biochem.* 236, 107–113 (1996).
41. Staden, R. The Staden sequence analysis package. *Mol. Biotech.* 5, 233–241 (1996).
42. Sicheritz-Pontén, T. BioWish: a molecular biology command extension to Tcl/Tk. *Comput. Biosci.* 13, 621–622 (1997).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search *Mol. Biol.* 215, 403–410 (1990).
44. Low, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA gen genomic sequence. *Nucleic Acids Res.* 25, 955–964 (1997).
45. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of prog multiple alignment through sequence weighting, position-specific gap penalties and weight m choice. *Nucleic Acids Res.* 22, 4673–4680 (1994).
46. Galtier, N., Gouy, M. & Gautier, C. SeaView and Phylo_win, two graphic tools for sequence align and molecular phylogeny. *Comput. Appl. Biosci.* 12, 543–548 (1996).
47. Sicheritz-Pontén, T. & Andersson, S. G. E. GRS: a graphic tool for genome retrieval and analysis. *Microb. Comp. Genomics* 2, 123–139 (1997).

Acknowledgements. We thank C. Woese for discussions; M. Andersen for computer system supp B. Andersson, K. Andersson, I. Tamas, B. Canbäck, A. Jamal, H. Amiri and S. Jossan for technical and assistance. This work was supported by the Swedish Foundation for Strategic Research, the Natural Sciences Research Council, the Knut and Alice Wallenberg Foundation and the Euro Commission.

Correspondence and requests for materials should be addressed to C.G.K. (e-mail: chuck bmc.uu.se).

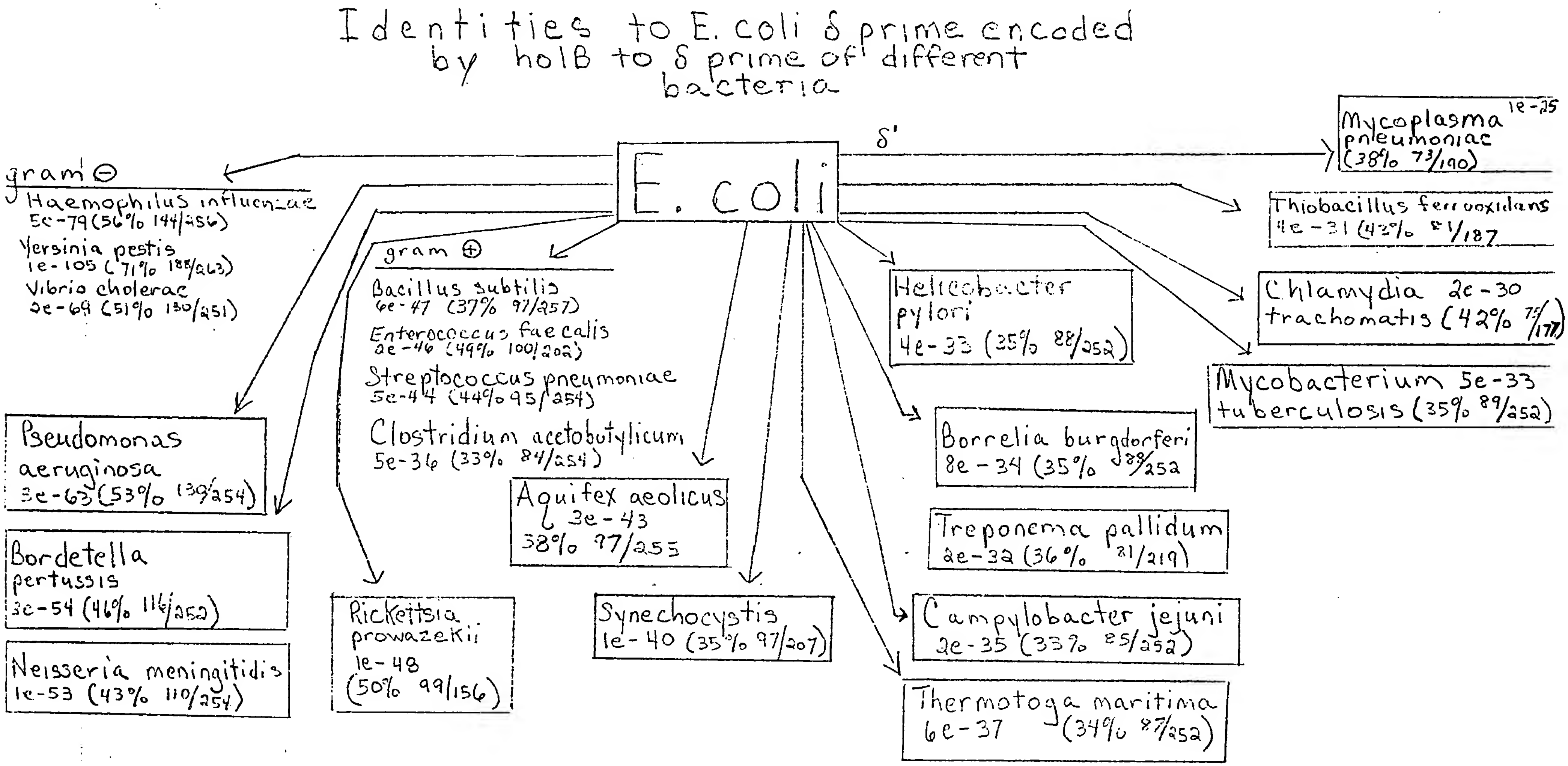

Identities to E. coli δ prime encoded by holB to δ prime of different bacteria
E. coli
δ'
gram ⊖
Haemophilus influenzae 5e-79 (56% 144/256)
Yersinia pestis 1e-105 (71% 188/263)
Vibrio cholerae 2e-69 (51% 130/251)
gram ⊕
Bacillus subtilis 6e-47 (37% 97/257)
Enterococcus faecalis 2e-46 (49% 100/202)
Streptococcus pneumoniae 5e-44 (44% 95/254)
Clostridium acetobutylicum 5e-36 (33% 84/254)
Pseudomonas aeruginosa 3e-63 (53% 139/254)
Bordetella pertussis 3e-54 (46% 116/252)
Neisseria meningitidis 1e-53 (43% 110/254)
Rickettsia prowazekii 1e-48 (50% 99/156)
Aquifex aeolicus 3e-43 38% 97/255
Synechocystis 1e-40 (35% 97/207)
Mycoplasma pneumoniae 1e-25 (38% 73/190)
Thiobacillus ferrooxidans 4e-31 (43% 81/187
Chlamydia trachomatis 2e-30 (42% 75/177)
Mycobacterium tuberculosis 5e-33 (35% 89/252)
Helicobacter pylori 4e-33 (35% 88/252)
Borrelia burgdorferi 8e-34 (35% 88/252
Treponema pallidum 2e-32 (36% 81/219)
Campylobacter jejuni 2e-35 (33% 85/252)
Thermotoga maritima 6e-37 (34% 87/252)

Figure 1

NCBI **BLAST Search Results** BLAST Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

You have searched a database generously provided by the Institute for Genomic Research (TIGR). Their Policy on Early Data Release is:

The Institute for Genomic Research (TIGR) releases data very rapidly to ensure that our scientific colleagues have access to information that may assist them in the search for genes and their biological function. Data releases do not constitute scientific publication, but rather provide investigators with information that may "jump-start" biological experimentation. Users of this information are encouraged to share their results with TIGR in order to improve annotation of the sequence data. Data or information may contain errors or be incomplete and should be regarded as preliminary.

TIGR asks that you acknowledge the source of information obtained from this site in any publication by including the following sentence in both the Materials and Methods and Acknowledgement sections: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" Also include the following text in the Acknowledgements, if applicable: "Sequencing of [organism name] was accomplished with support from [funding agency]." The name of the funding agency for each TIGR project can be found at http://www.tigr.org/tdb/mdb/mdb.html

Similarly, if you display this data or any information derived from it on a Web page, we ask that you prominently display the following notice on that webpage: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" We request that you notify us of your electronic presentation by sending email to www@tigr.org.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query=
(334 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

	Score (bits)	E Value
Sequences producing significant alignments:		

gb|U00096|ECOLI Escherichia coli K-12 MG1655 complete genome	614	e-175
gnl|Sanger|S.typhi_Contig369 Salmonella typhi unfinished fragmen...	490	e-138
gnl|Sanger|Y.pesits_Contig315 Yersinia pestis unfinished fragmen...	284	2e-76
gnl|CBCUMN|Pmultocida.990513.Contig500 Pasteurella multocida PM7...	175	2e-43
gnl|CBCUMN|PMultocida.990407.Contig485 Pasteurella multocida PM7...	174	3e-43
gb|L42023|L42023 Haemophilus influenzae Rd complete genome	153	6e-37
gnl|CBCUMN|F8P5 Pasteurella multocida PM70 unfinished fragment o...	141	4e-33
gnl|TIGR|V.cholerae_asm894 Vibrio cholerae unfinished fragment o...	123	6e-28
gnl|PAGP|Paeruginosa_Contig50 Pseudomonas aeruginosa unfinished ...	115	2e-25
gnl|OUACGT|A.actin_Contig398 Actinobacillus actinomycetemcomitan...	115	2e-25
gnl|TIGR|S.putrefaciens_gsp_271 Shewanella putrefaciens unfinish...	95	3e-19
gnl|Sanger|N.mening_Contig4 Neisseria meningitidis serogroup A u...	89	2e-17
gnl|OUACGT|Ngon_Contig191 Neisseria gonorrhoeae unfinished fragm...	87	6e-17
gnl|TIGR|D.radiodurans_8842 Deinococcus radiodurans unfinished f...	76	1e-13
gnl|Sanger|B.pertussis_Contig654 Bordetella pertussis unfinished...	73	1e-12
gnl|OUACGT|Ngon_Contig223 Neisseria gonorrhoeae unfinished fragm...	73	1e-12
gnl|Sanger|N.mening_Contig3 Neisseria meningitidis serogroup A u...	73	1e-12
emb|AL123456|MTBH37RV Mycobacterium tuberculosis H37Rv complete ...	69	2e-11
gnl|TIGR|gmt3711 Mycobacterium tuberculosis unfinished fragment ...	69	2e-11
gnl|Sanger_1765|mbovis_Contig1041.0 Mycobacterium bovis unfinish...	69	2e-11
gb|AE000657|AE000657 Aquifex aeolicus complete genome	68	6e-11
gnl|Sanger|B.pertussis_Contig889 Bordetella pertussis unfinished...	64	5e-10
gnl|TIGR|t_ferrooxidans_1986 Thiobacillus ferrooxidans unfinishe...	64	5e-10
gnl|TIGR|C.tepidum_gct_9 Chlorobium tepidum unfinished fragment ...	64	7e-10
gnl|TIGR|gef_6277 Enterococcus faecalis unfinished fragment of c...	63	2e-09
gnl|Sanger|Y.pesits_Contig790 Yersinia pestis unfinished fragmen...	63	2e-09
gnl|TIGR|C.trachomatis_ct_97 Chlamydia trachomatis MOPN unfinish...	62	2e-09
gnl|PAGP|Paeruginosa_Contig53 Pseudomonas aeruginosa unfinished ...	62	3e-09
gnl|TIGR|N.meningitidis_GNMCF18R Neisseria meningitidis MC58 unf...	61	6e-09
gnl|TIGR|T.maritima_tm_26 Thermotoga maritima unfinished fragmen...	61	6e-09
gnl|TIGR|P.gingivalis_1194 Porphyromonas gingivalis W83 unfinish...	60	1e-08
gnl|Sanger|S.typhi_Contig376 Salmonella typhi unfinished fragmen...	59	2e-08
gnl|OUACGT|Spyogenes_Contig243 Streptococcus pyogenes unfinished...	59	2e-08
gnl|TIGR|S.putrefaciens_gsp_387 Shewanella putrefaciens unfinish...	59	2e-08
gnl|TIGR|M.avium_5593 Mycobacterium avium unfinished fragment of...	58	4e-08
emb|AL009126|BSUB Bacillus subtilis complete genome	58	4e-08
gnl|UOKNOR|S.mutans_Contig840 Streptococcus mutans unfinished fr...	58	4e-08
gb|AE001273|AE001273 Chlamydia trachomatis complete genome	58	5e-08
gnl|TIGR|C.crescentus_gcc_764 Caulobacter crescentus unfinished ...	57	9e-08
gnl|Sanger|campylo_Cj.seq Campylobacter jejuni NCTC 11168 unfini...	57	1e-07
gnl|OUACGT|A.actin_Contig753 Actinobacillus actinomycetemcomitan...	56	2e-07
gnl|TIGR|gmt3732 Mycobacterium tuberculosis unfinished fragment ...	56	2e-07
gnl|GTC|C.aceto_AE001437 Clostridium acetobutylicum, WORKING DRA...	55	3e-07
gnl|TIGR|S.pneumoniae_sp_36 Streptococcus pneumoniae unfinished ...	55	4e-07
gnl|TIGR|V.cholerae_asm864 Vibrio cholerae unfinished fragment o...	54	6e-07
gnl|TIGR|P.gingivalis_1209 Porphyromonas gingivalis W83 unfinish...	54	8e-07
gnl|Sanger_1765|mbovis_Contig454.1 Mycobacterium bovis unfinishe...	53	1e-06
gb|AE000520|AE000520 Treponema pallidum complete genome	53	1e-06
gb|AE000511|HPYL Helicobacter pylori 26695 complete genome	52	4e-06
gnl|TIGR|S.aureus_2202 Staphylococcus aureus COL unfinished frag...	51	5e-06
gnl|OUACGT|S.aureus_Contig1164 Staphylococcus aureus unfinished ...	51	5e-06
gb|AE001439|AE001439 Helicobacter pylori, strain J99 complete ge...	50	1e-05
gnl|TIGR|N.meningitidis_GNMAB03R Neisseria meningitidis MC58 unf...	50	1e-05
gnl|TIGR|S.pneumoniae_sp_68 Streptococcus pneumoniae unfinished ...	49	2e-05
gnl|TIGR|C.tepidum_gct_35 Chlorobium tepidum unfinished fragment...	48	4e-05
gnl|TIGR|t_ferrooxidans_64 Thiobacillus ferrooxidans unfinished ...	48	4e-05
gnl|OUACGT|Ngon_Contig196 Neisseria gonorrhoeae unfinished fragm...	47	7e-05
gnl|TIGR|D.radiodurans_8813 Deinococcus radiodurans unfinished f...	47	7e-05
gb|AE000783|AE000783 Borrelia burgdorferi complete genome	47	7e-05
gnl|TIGR|t_ferrooxidans_1967 Thiobacillus ferrooxidans unfinishe...	46	2e-04
gnl|TIGR|gef_6250 Enterococcus faecalis unfinished fragment of c...	45	3e-04

```
emb|AJ235269|RPXX0 Rickettsia prowazekii strain Madrid E, comple...   42  0.003
gnl|OUACGT|Spyogenes_Contig260 Streptococcus pyogenes unfinished...   42  0.003
gnl|OUACGT|Ngon_Contig166 Neisseria gonorrhoeae unfinished fragm...   41  0.004
gnl|UOKNOR|S.mutans_Contig762 Streptococcus mutans unfinished fr...   41  0.007
gnl|TIGR|C.trachomatis_ct_26 Chlamydia trachomatis MOPN unfinish...   39  0.028
gnl|TIGR|S.aureus_2184 Staphylococcus aureus COL unfinished frag...   38  0.037
gnl|CBCUMN|Pmultocida.990513.Contig705 Pasteurella multocida PM7...   36  0.19
gb|AB001339|SYNECHO Synechocystis PCC6803 complete genome             36  0.25
gnl|TIGR|M.avium_5418 Mycobacterium avium unfinished fragment of...   32  2.1
gnl|TIGR|C.crescentus_gcc_2104 Caulobacter crescentus unfinished...   32  2.8
gnl|TIGR|V.cholerae_asm959 Vibrio cholerae unfinished fragment o...   31  8.3

 gb|U00096|ECOLI Escherichia coli K-12 MG1655 complete genome
           Length = 4639221

 Score =  614 bits (1566), Expect = e-175
 Identities = 296/334 (88%), Positives = 296/334 (88%)
 Frame = +3

Query: 1       MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG 60
               MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG
Sbjct: 1154985 MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG 1155164

Query: 61      HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXX 120
               HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWV
Sbjct: 1155165 HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALL 1155344

Query: 121     XXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREV 180
                           EEPPAETWFFLATREPERLLATLRSRCRLHYLA PPEQYAVTWLSREV
Sbjct: 1155345 TDAAANALLKTLEEPPAETWFFLATREPERLLATLRSRCRLHYLAPPPEQYAVTWLSREV 1155524

Query: 181     TMSQDXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQ 240
               TMSQD                   FQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQ
Sbjct: 1155525 TMSQDALLAALRLSAGSPGAALALFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQ 1155704

Query: 241     APARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQL 300
               APARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQL
Sbjct: 1155705 APARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQL 1155884

Query: 301     MSVTGINRELLITDLLLRIEHYLQPGVVLPVPHL 334
               MSVTGINRELLITDLLLRIEHYLQPGVVLPVPHL
Sbjct: 1155885 MSVTGINRELLITDLLLRIEHYLQPGVVLPVPHL 1155986


 Score = 57.4 bits (136), Expect = 7e-08
 Identities = 37/144 (25%), Positives = 59/144 (40%)
 Frame = +3

Query: 21     GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
              GR HHA L     G+G  ++   L++ L C+       CG C  C+ ++ G   D   +
Sbjct: 491418 GRIHHAYLFSGTRGVGKTSIARLLAKGLNCETGITATPCGVCDNCREIEQGRFVDLIEI- 491594

Query: 81     PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETW 140
                   +   V+  R++ + +      G  KV  +                  EEPP
Sbjct: 491595 --DAASRTKVEDTRDLLDNVQYAPARGRFKVYLIDEVHMLSRHSFNALLKTLEEPPEHVK 491768

Query: 141    FFLATREPERLLATLRSRCRLHYL 164
              F LAT +P++L  T+ SRC   +L
Sbjct: 491769 FLLATTDPQKLPVTILSRCLQFHL 491840
```

```
gnl|Sanger|S.typhi_Contig369 Salmonella typhi unfinished fragment of complete genome
          Length = 5674

 Score =  490 bits (1248), Expect = e-138
 Identities = 229/334 (68%), Positives = 262/334 (77%)
 Frame = -1

Query: 1    MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG 60
             M+WYPWLRP +EKLV SYQAGRGHHALLIQALPGMGD+AL YALSRYLLCQQP+GHKSCG
Sbjct: 2329 MKWYPWLRPAYEKLVESYQAGRGHHALLIQALPGMGDEALCYALSRYLLCQQPEGHKSCG 2150

Query: 61   HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXX 120
             HCRGCQLMQAGTHPDYYTL P+KGK++LGVDAVREV+EKL EH+RLGGAKVVW+
Sbjct: 2149 HCRGCQLMQAGTHPDYYTLTPDKGKSSLGVDAVREVSEKLYEHSRLGGAKVVWIADAALL 1970

Query: 121  XXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREV 180
                         EEPP +TWFFLA+ EP RLLATLRSRCRLH+LA P E YA++WLSREV
Sbjct: 1969 TDAAANALLKTLEEPPEQTWFFLASPEPARLLATLRSRCRLHHLAPPSESYAMSWLSREV 1790

Query: 181  TMSQDXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQ 240
             T SQ+                   Q + W  RE LCQAL  S+ +GDWY+LL ALNHEQ
Sbjct: 1789 TASQEALLTALRLNAGSPGAALALLQSERWAQREALCQALMDSLHTGDWYALLTALNHEQ 1610

Query: 241  APARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQL 300
             APARLHWLATLL+DALKR HGA+ +TNVD   +VA LA  LSP+R+QAIL DVCH R+QL
Sbjct: 1609 APARLHWLATLLVDALKRQHGASYLTNVDADAVVAALAGPLSPARIQAILNDVCHCRDQL 1430

Query: 301  MSVTGINRELLITDLLLRIEHYLQPGVVLPVPHL 334
             + VTG+NREL++TDL+LRIEHYLQPG +L VPHL
Sbjct: 1429 LHVTGLNRELVLTDLILRIEHYLQPGTLLXVPHL 1328


gnl|Sanger|Y.pesits_Contig315 Yersinia pestis unfinished fragment of complete genome
          Length = 20197

 Score =  284 bits (720), Expect = 2e-76
 Identities = 147/334 (44%), Positives = 192/334 (57%), Gaps = 6/334 (1%)
 Frame = -1

Query: 1     MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG 60
              M WYPWL    + +LV  +   GRGHHALL+ +LPG G+DALIYALSR+L+CQQ QG KSCG
Sbjct: 15274 MNWYPWLNAPYRQLVGQHSTGRGHHALLLHSLPGNGEDALIYALSRWLMCQQRQGEKSCG 15095

Query: 61    HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXX 120
               C  C+LM AG HPD+Y L PEKGK+++GV+ VR++ +KL  HA+ GGAKVVW+
Sbjct: 15094 ECHSCRLMLAGNHPDWYVLTPEKGKSSIGVELVRQLIDKLYSHAQQGGAKVVWLPHAEVL 14915

Query: 121   XXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLS--- 177
                          EEPP +T+F L    +P  LLATLRSRC    YLA P      + WL+
Sbjct: 14914 TDAAANALLKTLEEPPEKTYFLLDCHQPASLLATLRSRCFYWYLACPDTAICLQWLNLQW 14735

Query: 178   --REVTMSQDXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAA 235
                R++ +                      Q + W  R  LC  L  ++    D  SLL
Sbjct: 14734 RKRQIPVEPVAMLAALKLSEGAPLAAERLLQPERWSIRSALCSGLREALNRSDLLSLLPQ 14555

Query: 236   LNHEQAPARLHWLATLLMDALKRHHGAAQ-VTNVDVPGLVAELANHLSPSRLQAILGDVC 294
              LNH+ A  RL WL++LL+DALK   GA +   N D   LV +LA+  +    L  +   +
Sbjct: 14554 LNHDDAAERLQWLSSLLLDALKWQQGAGEFAVNQDQLPLVQQLAHIAATPVLLQLAKQLA 14375

Query: 295   HIREQLMSVTGINRELLITDLLLRIEHYLQPGVVLPVPHL 334
              H R QL+SV G+NRELL+T+ LL  E  L  G    +P L
```

```
Sbjct: 14374 HCRHQLLSVVGVNRELLLTEQLLSWETALSTGTYSTLPSL 14255


 gnl|CBCUMN|Pmultocida.990513.Contig500 Pasteurella multocida PM70 unfinished fragment of
            Length = 1241

 Score =  175 bits (439), Expect = 2e-43
 Identities = 102/319 (31%), Positives = 151/319 (46%), Gaps = 4/319 (1%)
 Frame = -1

Query: 1    MRWYPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCG 60
            M  YPWL P +++ + ++Q G GHHALL QA  G+   + L++AL  +L+CQQPQ  + C
Sbjct: 1196 MTLYPWLLPYYQQRIDAFQQGHGHHALLFQAEQGLSTEQLLFALGHWLICQQPQNQQPCQ 1017

Query: 61   HCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXX 120
             C  C L QA THPD YTL P + K+ +GVD VREV EK+N+HA+ GG K+++V
Sbjct: 1016 QCHHCHLFQAQTHPDIYTLTPIENKD-IGVDQVREVNEKINQHAQQGGNKIIYVLGVSRL 840

Query: 121  XXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREV 180
                        EEP   T+F L T   + ++ T+ SRC+   LA P E  A+ WL ++
Sbjct: 839  TEAAANAMLKTLEEPRPNTYFLLYTEASDSVMPTIYSRCQTQKLALPAETSAIAWLQQQT 660

Query: 181  TMSQDXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQ 240
            T                          Q D  +R     +                LL    +
Sbjct: 659  TQEIAAIQTALRISYGRPLHALTVLQQDLLEKRREFLRQFWLFYRKRSPLELLPFFDKAI 480

Query: 241  APARLHWLATLLMDALKRHHGAAQVTN----VDVPGLVAELANHLSPSRLQAILGDVCHI 296
               +L WL    L DALK       Q+ +      D+   V +L+    S    L        +  +
Sbjct: 479  LLHQLDWLLAFLSDALK---AKLQIKSDWLCQDLAAGVLQLSQQQSAQALLHATQIIQKV 309

Query: 297  REQLMSVTGINRELLITDLLLRI 319
            R  L  +  +N+EL++ D L ++
Sbjct: 308  RTDLTQINAVNQELILLDGLTQL 240


 gnl|CBCUMN|PMultocida.990407.Contig485 Pasteurella multocida PM70 unfinished fragment of
            Length = 1370

 Score =  174 bits (437), Expect = 3e-43
 Identities = 101/316 (31%), Positives = 150/316 (46%), Gaps = 4/316 (1%)
 Frame = -3

Query: 4    YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
            YPWL P +++ + ++Q G GHHALL QA  G+   + L++AL  +L+CQQPQ  + C  C
Sbjct: 1218 YPWLLPYYQQRIDAFQQGHGHHALLFQAEQGLSTEQLLFALGHWLICQQPQNQQPCQQCH 1039

Query: 64   GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
             C L QA THPD YTL P + K+ +GVD VREV EK+N+HA+ GG K+++V
Sbjct: 1038 HCHLFQAQTHPDIYTLTPIENKD-IGVDQVREVNEKINQHAQQGGNKIIYVLGVSRLTEA 862

Query: 124  XXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMS 183
                     EEP   T+F L T   + ++ T+ SRC+   LA P E  A+ WL ++ T
Sbjct: 861  AANAMLKTLEEPRPNTYFLLYTEASDSVMPTIYSRCQTQKLALPAETSAIAWLQQQTTQE 682

Query: 184  QDXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQAPA 243
                                 Q D  +R     +                LL    +
Sbjct: 681  IAAIQTALRISYGRPLHALTVLQQDLLEKRREFLRQFWLFYRKRSPLELLPFFDKAILLH 502

Query: 244  RLHWLATLLMDALKRHHGAAQVTN----VDVPGLVAELANHLSPSRLQAILGDVCHIREQ 299
            +L WL    L DALK       Q+ +      D+   V +L+    S    L        +  +R
Sbjct: 501  QLDWLLAFLSDALK---AKLQIKSDWLCQDLAAGVLQLSQQQSAQALLHATQIIQKVRTD 331
```

```
Query: 300   LMSVTGINRELLITDLLLRI 319
             L  +  +N+EL++ D L ++
Sbjct: 330   LTQINAVNQELILLDGLTQL 271
```

gb|L42023|L42023 Haemophilus influenzae Rd complete genome
Length = 1830138

Score = 153 bits (384), Expect = 6e-37
Identities = 97/316 (30%), Positives = 150/316 (46%), Gaps = 7/316 (2%)
Frame = -2

```
Query: 4      YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
              YPWL P + ++   ++  G GHHA+LI+A  G+G ++L  AL++ ++C   QG K CG C
Sbjct: 477329 YPWLMPIYHQIAQTFDEGLGHHAVLIKADSGLGVESLFNALAQKIMCVA-QGDKPCGQCH 477153

Query: 64     GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
               C LMQA +HPDY+ L+P  GK+ +GVD VR++ E + +HA+  G KVV+V
Sbjct: 477152 SCHLMQAHSHPDYHELSPINGKD-IGVDQVRDINEMVAQHAQQNGNKVVYVQGAERLTEA 476976

Query: 124    XXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMS 183
                       EEP   T+F L       LLAT+ SRC++  L+ P E+ A  WL  +  +
Sbjct: 476975 AANALLKTLEEPRPNTYFLLQADSSASLLATIYSRCQVWNLSVPNEEIAFEWLKSKSAVE 476796

Query: 184    QDXXXXXXXXXXXXXXXXXXXXFQGDNWQARETLCQALAYSVPSGDWYSLLAALNHEQAPA 243
                                  Q    + R+    +                LL    + E+
Sbjct: 476795 NQEILTALAMNLGRPLLALETLQEGFIEQRKNFLRQFWVFYRRRSPLELLPLFDKERYVQ 476616

Query: 244    RLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSP-SRLQAILG------DVCHI 296
              ++ W+    L D LK           +D     VA+L   +   S  Q  LG        +  +
Sbjct: 476615 QVDWILAFLSDCLKHK------LEIDSHRQVADLGRGIEQFSDEQTALGLLQAIKIMQKV 476454

Query: 297    REQLMSVTGINRELLITDLLLRI 319
              R  L+++ G+N EL++ D L R+
Sbjct: 476453 RSDLLTINGVNVELMLLDGLTRL 476385
```

Score = 56.6 bits (134), Expect = 1e-07
Identities = 36/143 (25%), Positives = 57/143 (39%)
Frame = -2

```
Query: 22      RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
               R HHA L     G+G  ++     ++ L C          CG C  C+ ++ G    D    +
Sbjct: 1299740 RLHHAYLFSGTRGVGKTSIARLFAKGLNCVHGVTATPCGECENCKAIEQGNFIDLIEI-- 1299567

Query: 82      EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWF 141
                  +    V+   RE+ + +        +G   KV   +                  EEPP     F
Sbjct: 1299566 -DAASRTKVEDTRELLDNVQYKPVVGRFKVYLIDEVHMLSRHSFNALLKTLEEPPEYVKF 1299390

Query: 142     FLATREPERLLATLRSRCRLHYL 164
                LAT +P++L  T+ SRC   +L
Sbjct: 1299389 LLATTDPQKLPVTILSRCLQFHL 1299321
```

gnl|CBCUMN|F8P5 Pasteurella multocida PM70 unfinished fragment of complete genome
Length = 550

Score = 141 bits (351), Expect = 4e-33
Identities = 64/149 (42%), Positives = 90/149 (59%)
Frame = +3

```
Query: 4   YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
           YPWL P  +++ + ++Q G GHHALL QA  G+G + L++AL  +L+CQQPQ  + C  C
Sbjct: 9   YPWLLPYYQQRIDAFQQGHGHHALLFQAEQGLGTEQLLFALGHWLICQQPQNQQPCQQCH 188

Query: 64  GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
             C L QA THPD YTL P + K+ +GVD VREV EK+N+HA+ GG K+++V
Sbjct: 189 HCHLFQAQTHPDIYTLTPIENKD-IGVDQVREVNEKINQHAQQGGNKIIYVLGVSRLTEA 365

Query: 124 XXXXXXXXXEEPPAETWFFLATREPERLL 152
                    EEP   T+F L T   + ++
Sbjct: 366 AANAMLKTLEEPRPNTYFLLYTEASDSVM 452
```

gnl|TIGR|V.cholerae_asm894 Vibrio cholerae unfinished fragment of complete genome
Length = 19711

Score = 123 bits (307), Expect = 6e-28
Identities = 90/313 (28%), Positives = 136/313 (42%), Gaps = 3/313 (0%)
Frame = -1

```
Query: 4     YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
             YPWL P ++   A   AG+   A LIQA  G+G ++L+   ++R L+C   Q  + CG C
Sbjct: 18034 YPWLVPVWQPWQAGLAAGKISSATLIQASEGVGVESLVELMARTLMCTSSQS-EPCGFCH 17858

Query: 64    GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
               C LMQ+G HPD++ + PEK    ++ V+ +R++      E ++L G +++ +
Sbjct: 17857 SCGLMQSGNHPDFHVVKPEKIGKSITVEQIRQMNRIAQESSQLSGYRLIVIEPADAMNES 17678

Query: 124   XXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMS 183
                      EEP     F L T   + LL T+ SRC+    L  P     V WL  +    +
Sbjct: 17677 SANALLKTLEEPAPNCLFILVTSRIKHLLPTIVSRCQRLVLPAPTTALVVEWLKGQGITT 17498

Query: 184   QDXXXXXXXXXXXXXXXXXXXFQGDNWQARET-LCQALAYSVPSGDWYSLL--AALNHEQ 240
                                   + +     E+ L  AL      SGD  + L   AL
Sbjct: 17497 PAYALHLCADSPLKTRAFMLEGGAEKYHELESQLMNAL-----SGDVNAQLKCIALIDAD 17333

Query: 241   APARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHLSPSRLQAILGDVCHIREQL 300
                 L+W+   +L DA K H G  Q        P   A LA    + S+L       +  + EQL
Sbjct: 17332 LTTHLYWVWCVLTDAQKIHFGVQQDY---YPPASAALAGRFTYSKLHVQTASLERLMEQL 17162

Query: 301   MSVTGINRELLITDLL 316
                 +G+N ELL+    L
Sbjct: 17161 NQFSGLNTELLLLQWL 17114
```

gnl|PAGP|Paeruginosa_Contig50 Pseudomonas aeruginosa unfinished fragment of complete ger
Length = 798876

Score = 115 bits (286), Expect = 2e-25
Identities = 84/323 (26%), Positives = 139/323 (43%), Gaps = 11/323 (3%)
Frame = +2

```
Query: 4      YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCR 63
              YPW +   + +L     Q      HA L+      G+G  AL      +  LLCQ+P     +CG C+
Sbjct: 521618 YPWQQALWSQLGGRAQHA---HAYLLYGPAGIGKRALAEHWAAQLLCQRPAAAGACGECK 521788

Query: 64     GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
                CQL+ AGTHPDY+ L PE+ +   + VD VR++     + + A+LGG KVV +
Sbjct: 521789 ACQLLAAGTHPDYFVLEPEEAEKPIRVDQVRDLVGFVVQTAQLGGRKVVLLEPAEAMNVN 521968
```

```
Query: 124    XXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMS 183
                       EEP  +T   L + +P RLL T++SRC       P    ++ WL+R +
Sbjct: 521969 AANALLKSLEEPSGDTVLLLISHQPSRLLPTIKSRCVQQACPLPGAAASLEWLARALPDE 522148

Query: 184    QDXXXXXXXXXXXXXXXXXXXFQGDNWQ---------ARETLCQALAYSVPSGDWYSLLA 234
                                 G   +          ++ L Q +A S  +   W
Sbjct: 522149 PAEALEELLALSGGSPLTAQRLHGQGVREQRAQVVEGVKKLLKQQIAASPLAESW----- 522313

Query: 235    ALNHEQAPARLHWLATLLMDALKRH--HGAAQVTNVDVPGLVAELANHLSPSRLQAILGD 292
                N   P   W    + L+   H    +   D+  ++   L +    +++ A+
Sbjct: 522314 --NSVPLPLLFDWFCDWTLGILRYQLTHDEEGLGLADMRKVIQYLGDKSGQAKVLAMQDW 522487

Query: 293    VCHIREQLMSVTGINRELLITDLLLRIEHYLQPG 326
              +   R+++++   +NR LL+  LL++       PG
Sbjct: 522488 LLQQRQKVLNKANLNRVLLLEALLVQWASLPGPG 522589
```

gnl|OUACGT|A.actin_Contig398 Actinobacillus actinomycetemcomitans unfinished fragment of genome
Length = 1469

Score = 115 bits (285), Expect = 2e-25
Identities = 88/293 (30%), Positives = 136/293 (46%), Gaps = 4/293 (1%)
Frame = -1

```
Query: 27   LLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKGKN 86
            LLI+A  G+G + L   L++ L+C   P+   + CG C   C LMQA +HPD+   +AP + K+
Sbjct: 1469 LLIRADEGLGAEQLCRLLAQRLMCLTPKSAEPCGECHACHLMQANSHPDFQHIAPIENKD 1290

Query: 87   TLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLATR 146
             +GVD +R + E+ ++HA+  G KV++++                EEP   T+F L
Sbjct: 1289 -IGVDQIRAMNEQASQHAQQNGNKVIYIEQAHRLTESAANAILKTLEEPRPNTYFILQND 1113

Query: 147  EPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTM-SQDXXXXXXXXXXXXXXXXXXXF 205
              + LL T+ SRC++   L  P    A+ WL  + ++ + +
Sbjct: 1112 MQKALLPTIYSRCQVWNLLPPATDTALHWLQAQTSVETPEILTALLVNYGRPLLALAMLT 933

Query: 206  QGDNWQARETLCQA-LAYSVPSGDWYSLLAALNHEQAPARLHWLATLLMDALKRHHGAAQ 264
            Q    Q RE L Q  L Y   S     LL   N E    +L WL   L D+LK +  A Q
Sbjct: 932  QHLPEQRREFLRQFWLFYRRRSP--LELLPFFNKEILLQQLDWLLAFLSDSLK-NKLAIQ 762

Query: 265  VTNV--DVPGLVAELANHLSPSRLQAILGDVCHIREQLMSVTGINRELLITDLLLRI 319
               +  D+   V + +  LS   L        V +R  L +  +N+EL++ D L R+
Sbjct: 761  ENWICRDIERGVIQFSQGLSAPALLKATQIVGKVRSDLAANNALNQELILLDGLTRL 591
```

gnl|TIGR|S.putrefaciens_gsp_271 Shewanella putrefaciens unfinished fragment of complete
Length = 11991

Score = 95.2 bits (233), Expect = 3e-19
Identities = 51/181 (28%), Positives = 81/181 (44%)
Frame = +3

```
Query: 5    PWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG 64
            PWL     +  +   Q  +  HA L+      G + L   ++R  +C QP      CG C+
Sbjct: 1842 PWLDVPRQAFLTQLQTQKVPHAQLVGIDSAYGGELLSVFMARAAMCSQPTHTGGCGFCKS 2021

Query: 65   CQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXX 124
            CQL  AG HPD+Y +   E    + + VD +RE+  +L+  A+  G +V   +
Sbjct: 2022 CQLFDAGNHPDFYQI--EADGHQIKVDQIRELCSRLSATAQQSGRRVAIIHHSERLNSAS 2195
```

```
Query: 125  XXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMSQ 184
                    EEP  +T   L +  P RL+AT+ SRC+      P +     WL ++  + +
Sbjct: 2196 ANALLKTLEEPGKDTLLLLHSDTPARLMATISSRCQRLPFVAPSKTLIKNWLIQQCQIQE 2375

Query: 185  D 185
            D
Sbjct: 2376 D 2378


 gnl|Sanger|N.mening_Contig4 Neisseria meningitidis serogroup A unfinished fragment of cc
            Length = 236507

 Score = 88.9 bits (217), Expect = 2e-17
 Identities = 53/173 (30%), Positives = 86/173 (49%), Gaps = 8/173 (4%)
 Frame = -3

Query: 4      YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQP-QGHKSCGHC 62
              YPW +  + + +A +   R + A L     G G  A     ++ LLC++P  G+  CG C
Sbjct: 209151 YPWHQEQWRQ-IAEHWTSRPN-AWLFVGKKGTGKTAFARFAAKALLCEKPVTGNVPCGEC 208978

Query: 63     RGCQLMQAGTHPDYYTLAP-----EKGKNTLGV--DAVREVTEKLNEHARLGGAKVVWVX 115
                C L + G+HPD+Y + P     E G+  L +  DAVRE+ + +   +  GG +V+ +
Sbjct: 208977 ASCHLFEQGSHPDFYEITPLTDERENGRKLLQIKIDAVREIIDNVYLTSVRGGLRVILIH 208798

Query: 116    XXXXXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTW 175
                               EEPP + F L +   +++L T++SRCR   L  P  + A  +
Sbjct: 208797 PAESMNVQAANSLLKVLEEPPPQVVFLLVSHAADKVLPTIKSRCRKMVLPAPSHEEASAY 208618

Query: 176    L 176
              L
Sbjct: 208617 L 208615


 gnl|OUACGT|Ngon_Contig191 Neisseria gonorrhoeae unfinished fragment of complete genome
          Length = 20169

 Score = 87.4 bits (213), Expect = 6e-17
 Identities = 54/173 (31%), Positives = 84/173 (48%), Gaps = 8/173 (4%)
 Frame = -1

Query: 4    YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQ-GHKSCGHC 62
            YPW +  + + +A +   R + A L     G G  A     ++ LLC+ P  G K CG C
Sbjct: 4188 YPWHQEQWRQ-IAEHWTSRPN-AWLFVGKKGTGKTAFARFAAKALLCETPAPGCKPCGEC 4015

Query: 63   RGCQLMQAGTHPDYYTLAP-----EKGKNTLGV--DAVREVTEKLNEHARLGGAKVVWVX 115
              C L   G+HPD+Y + P     E G+  L +  DAVRE+ + +   +  GG +V+ +
Sbjct: 4014 MSCHLFGRGSHPDFYEITPLADEPENGRKLLRIKIDAVREIIDNVYLTSVRGGLRVILIH 3835

Query: 116  XXXXXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTW 175
                             EEPP + F L +   +++L T++SRCR   L  P    A+ +
Sbjct: 3834 PAESMNVQAANSLLKVLEEPPPQVVFLLVSHAADKVLPTIKSRCRKMVLPAPSHGEALAY 3655

Query: 176  L 176
            L
Sbjct: 3654 L 3652


 gnl|TIGR|D.radiodurans_8842 Deinococcus radiodurans unfinished fragment of complete genc
            Length = 18340

 Score = 76.5 bits (185), Expect = 1e-13
```

```
 Identities = 47/150 (31%), Positives = 67/150 (44%)
 Frame = +2

Query: 14    LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
              L  + + GR  HA L     G+G        ++     C  P    K CG C  C  ++AG+H
Sbjct: 13439 LRTALEQGRIGHAYLFSGPRGVGKTTTARLIAMTANCTGP-APKPCGECESCLAVRAGSH 13615

Query: 74    PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXE 133
              PD    +        +   VD VR++ EK+    A  GG K+   +                   E
Sbjct: 13616 PDVMEIDAASNNS---VDDVRDLREKVGLAAMRGGKKIYILDEAHMMSRAAFNALLKTLE 13786

Query: 134   EPPAETWFFLATREPERLLATLRSRCRLHY 163
              EPP    F LAT EPE+++ T+ SRC+ HY
Sbjct: 13787 EPPEHVIFILATTEPEKIIPTILSRCQ-HY 13873


 gnl|Sanger|B.pertussis_Contig654 Bordetella pertussis unfinished fragment of complete ge
          Length = 10062

 Score = 73.4 bits (177), Expect = 1e-12
 Identities = 55/178 (30%), Positives = 77/178 (42%), Gaps = 30/178 (16%)
 Frame = -2

Query: 2     RWYPWLRPDFEKLVASYQAGRGH--HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSC 59
              R+ PW       ++  S+ +GR     HA LI     G+G        A +   LLC+ P+     +C
Sbjct: 6023  RFLPWQT----EIARSWLSGRDRFAHAWLIHGNGGIGKLDFTAAAAASLLCESPRQGLAC 5856

Query: 60    GHCRGCQLMQAGTHPDYYTLAPEK----------------------------GKNTLGVD 91
              G C  C  + +G HPD    + PE                             + +D
Sbjct: 5855  GECAACAWVASGNHPDLRRIRPEAVALEEGADQTEGAEEAEAGSGGAAAKRAPSKDIRID 5676

Query: 92    AVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLATREPERL 151
               +R +    N     GG +V   +                   EEPPA T F L     P+RL
Sbjct: 5675  QIRALEPWFNTATHRGGWRVALLYPAHALNVISANALLKVLEEPPAHTVFLLVADAPDRL 5496

Query: 152   LATLRSRCRLHYLAGPPEQYAVTWLSRE 179
              L TL SRCR    L      A+ WL   +
Sbjct: 5495  LPTLVSRCRRLPLPTXSAGQALQWLGEQ 5412


 gnl|OUACGT|Ngon_Contig223 Neisseria gonorrhoeae unfinished fragment of complete genome
          Length = 90586

 Score = 73.0 bits (176), Expect = 1e-12
 Identities = 44/139 (31%), Positives = 62/139 (43%)
 Frame = -2

Query: 21    GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
              GR HHA L+     G+G     +     L++ L C+   Q   + CG C+ C   + AG + D    L
Sbjct: 72852 GRLHHAYLLTGTRGVGKTTIARILAKSLNCENAQHGEPCGVCQSCTQIDAGRYVD--LLE 72679

Query: 81    PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETW 140
               +    NT G+D +REV E        G  KV  +                   EEPP
Sbjct: 72678 IDAASNT-GIDNIREVLENAQYAPTAGKYKVYIIDEVHMLSKSAFNAMLKTLEEPPEHVK 72502

Query: 141   FFLATREPERLLATLRSRC 159
              F LAT +P ++   T+ SRC
Sbjct: 72501 FILATTDPHKVPVTVLSRC 72445


 gnl|Sanger|N.mening_Contig3 Neisseria meningitidis serogroup A unfinished fragment of cc
```

```
                Length = 291782

 Score = 73.0 bits (176), Expect = 1e-12
 Identities = 44/139 (31%), Positives = 62/139 (43%)
 Frame = +2

Query: 21     GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
              GR HHA L+    G+G    +    L++ L C+  Q   + CG C+ C   + AG + D    L
Sbjct: 180815 GRLHHAYLLTGTRGVGKTTIARILAKSLNCENAQHGEPCGVCQSCTQIDAGRYVD--LLE 180988

Query: 81     PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETW 140
               +    NT G+D +REV E          G   KV   +                        EEPP
Sbjct: 180989 IDAASNT-GIDNIREVLENAQYAPTAGKYKVYIIDEVHMLSKSAFNAMLKTLEEPPEHVK 181165

Query: 141    FFLATREPERLLATLRSRC 159
              F LAT +P ++   T+ SRC
Sbjct: 181166 FILATTDPHKVPVTVLSRC 181222


 Score = 41.4 bits (95), Expect = 0.004
 Identities = 16/37 (43%), Positives = 25/37 (67%)
 Frame = +2

Query: 4      YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDAL 40
              YPWL P + ++   ++   G GHHA+LI+A  G+G + L
Sbjct: 268937 YPWLMPIYHQIAQTFDEGLGHHAVLIKADAGLGVERL 269047


 emb|AL123456|MTBH37RV Mycobacterium tuberculosis H37Rv complete genome
                Length = 4411529

 Score = 69.1 bits (166), Expect = 2e-11
 Identities = 49/158 (31%), Positives = 68/158 (43%), Gaps = 5/158 (3%)
 Frame = -3

Query: 16      ASYQAGRGH---HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGT 72
               +++ AG G     HA L+    PG G          +   L C     G    CG CR C     AGT
Sbjct: 4082634 SAHSAGGGGTMTHAWLLTGPPGSGRSVAALCFAAALQCTSG-GEPGCGRCRACTTTLAGT 4082458

Query: 73      HPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXX 132
               H D     + PE      ++GVD +R + +           G   ++V +
Sbjct: 4082457 HADVRRVIPE--GLSIGVDEMRAIVQIAARRPTTGHWQIVVIEDADRLTEGAANALLKVV 4082284

Query: 133     EEPPAETWFFLA--TREPERLLATLRSRCRLHYLAGPPEQYAV 173
               EEPP   T F L     + +PE +   TLRSRCR H       P   +A+
Sbjct: 4082283 EEPPPSTVFLLCAPSVDPEDIAVTLRSRCR-HVALVTPSTHAI 4082158


 Score = 55.8 bits (132), Expect = 2e-07
 Identities = 44/150 (29%), Positives = 58/150 (38%)
 Frame = -2

Query: 14      LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
               L  +   AGR +HA L      G G   +     L+R L C Q         CG C   C  + A
Sbjct: 4166656 LSVALDAGRINHAYLFSGPRGCGKTSSARILARSLNCAQGPTANPCGVCESCVSL-APNA 4166480

Query: 74      PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXE 133
               P    +          +  GVD   RE+ ++                 +V   V                         E
Sbjct: 4166479 PGSIDVVELDAASHGGVDDTRELRDRAFYAPVQSRYRVFIVDEAHMVTTAGFNALLKIVE 4166300

Query: 134     EPPAETWFFLATREPERLLATLRSRCRLHY 163
```

```
                EPP    F  AT EPE++L T+RSR    HY
Sbjct: 4166299 EPPEHLIFIFATTEPEKVLPTIRSRTH-HY 4166213


 gnl|TIGR|gmt3711 Mycobacterium tuberculosis unfinished fragment of complete genome
            Length = 56385

 Score = 69.1 bits (166), Expect = 2e-11
 Identities = 49/158 (31%), Positives = 68/158 (43%), Gaps = 5/158 (3%)
 Frame = -3

Query: 16    ASYQAGRGH---HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGT 72
             +++ AG G    HA L+    PG G          +   L C     G    CG CR C     AGT
Sbjct: 44926 SAHSAGGGGTMTHAWLLTGPPGSGRSVAALCFAAALQCTSG-GEPGCGRCRACTTTLAGT 44750

Query: 73    HPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXX 132
             H D    + PE    ++GVD +R + +          G  ++V +
Sbjct: 44749 HADVRRVIPE--GLSIGVDEMRAIVQIAARRPTTGHWQIVVIEDADRLTEGAANALLKVV 44576

Query: 133   EEPPAETWFFLA--TREPERLLATLRSRCRLHYLAGPPEQYAV 173
             EEPP   T F L    + +PE +   TLRSRCR H       P   +A+
Sbjct: 44575 EEPPPSTVFLLCAPSVDPEDIAVTLRSRCR-HVALVTPSTHAI 44450


 gnl|Sanger_1765|mbovis_Contig1041.0 Mycobacterium bovis unfinished fragment of complete
            Length = 10794

 Score = 69.1 bits (166), Expect = 2e-11
 Identities = 49/158 (31%), Positives = 68/158 (43%), Gaps = 5/158 (3%)
 Frame = +3

Query: 16   ASYQAGRGH---HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGT 72
            +++ AG G    HA L+    PG G          +   L C     G    CG CR C     AGT
Sbjct: 4962 SAHSAGGGGTMTHAWLLTGPPGSGRSVAALCFAAALQCTSG-GEPGCGRCRACTTTLAGT 5138

Query: 73   HPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXX 132
            H D    + PE    ++GVD +R + +          G  ++V +
Sbjct: 5139 HADVRRVIPE--GLSIGVDEMRAIVQIAARRPTTGHWQIVVIEDADRLTEGAANALLKVV 5312

Query: 133  EEPPAETWFFLA--TREPERLLATLRSRCRLHYLAGPPEQYAV 173
            EEPP   T F L    + +PE +   TLRSRCR H       P   +A+
Sbjct: 5313 EEPPPSTVFLLCAPSVDPEDIAVTLRSRCR-HVALVTPSTHAI 5438


 gb|AE000657|AE000657 Aquifex aeolicus complete genome
              Length = 1551335

 Score = 67.5 bits (162), Expect = 6e-11
 Identities = 39/136 (28%), Positives = 58/136 (41%)
 Frame = +1

Query: 25      HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
                HA L       G+G      +      L++ L C+ P       + CG C     C+ +  G    PD    +
Sbjct: 1303996 HAYLFAGPRGVGKTTIARILAKALNCKNPSKGEPCGECENCREIDRGVFPDLIEMDAASN 1304175

Query: 85      KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                +     G+D VR + E +N      G  KV  +                       EEPP   T F L
Sbjct: 1304176 R---GIDDVRALKEAVNYKPIKGKYKVYIIDEAHMLTKEAFNALLKTLEEPPPRTVFVLC 1304346

Query: 145     TREPERLLATLRSRCR 160
                T E +++L T+ SRC+
```

```
Sbjct: 1304347 TTEYDKILPTILSRCQ 1304394


 Score = 43.0 bits (99), Expect = 0.001
 Identities = 35/132 (26%), Positives = 56/132 (41%), Gaps = 28/132 (21%)
 Frame = +3

Query: 27      LLIQALPGMGDDALIYALSRYLLCQQ--PQGHKSCGHCRGCQLMQA-------------- 70
                LL     G G       +   ++ +LC++   P G   SC  C+     ++
Sbjct: 1082652 LLFYGKEGSGKTKTAFEFAKGILCKENVPWGCGSCPSCKHVNELEEAFFKGEIEDFKVYK 1082831

Query: 71      ------------GTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXX 118
                            G HPD+   + P     + + ++ +REV          L    KV+ +
Sbjct: 1082832 DKDGKKHFVYLMGEHPDFVVIIPSG--HYIKIEQIREVKNFAYVKPALSRRKVIIIDDAH 1083005

Query: 119     XXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSR 158
                              EEPPA+T  F L T      +L T+ SR
Sbjct: 1083006 AMTSQAANALLKVLEEPPADTTFILTTNRRSAILPTILSR 1083125


 gnl|Sanger|B.pertussis_Contig889 Bordetella pertussis unfinished fragment of complete ge
          Length = 1034

 Score = 64.4 bits (154), Expect = 5e-10
 Identities = 41/138 (29%), Positives = 57/138 (40%)
 Frame = +2

Query: 22  RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
            R HHA L      G+G    L     L++ L C+       K CG CR C   + AG    DY   L
Sbjct: 626 RLHHAWLFTGTRGVGKTTLSRILAKSLNCENGITSKPCGQCRACTEIDAGRFVDYLELDA 805

Query: 82  EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWF 141
              +    GV+ + ++ E+         G  KV   +                  EEPP     F
Sbjct: 806 ASNR---GVEEMTQLLEQAVYAPGAGRFKVYMIDEVHMLTGHAFNAMLKTLEEPPPHVKF 976

Query: 142 FLATREPERLLATLRSRC 159
             LAT +P+ +   T+ SRC
Sbjct: 977 ILATTDPQIIPVTVLSRC 1030


 gnl|TIGR|t_ferrooxidans_1986 Thiobacillus ferrooxidans unfinished fragment of complete g
          Length = 733

 Score = 64.4 bits (154), Expect = 5e-10
 Identities = 46/149 (30%), Positives = 66/149 (43%), Gaps = 7/149 (4%)
 Frame = -3

Query: 28  LIQALPGMGDDALIYA-----LSRYLLCQQPQGHK-SCGHCRGCQLMQAGTHPDYYTLAP 81
            L QA+    G+     + A        L +   LC   P         CG CR C+L+   G HPD     + P
Sbjct: 542 LPQAMLAAGESGTLVAQYCDDLQQVALCFAPTAQGLPCGTCRSCRLLAEGNHPDLLMITP 363

Query: 82  EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWF 141
            E GK   + ++AVR    E L       ++       + + +                    EEP A
Sbjct: 362 ETGKR-ITIEAVRHANEFLAFTPQVSACRWLRIAPAEAMTAAAANALLKTLEEPAARAHI 186

Query: 142 FLATREPERLLATLRSRC-RLHYLAGPPEQYAVTWL 176
             L +    P +L+ T+RSR   RL +      P Q   V WL
Sbjct: 185 LLLSEHPSQLIPTIRSRLQRLPFPTMLPGQ-CVNWL 81


 gnl|TIGR|C.tepidum_gct_9 Chlorobium tepidum unfinished fragment of complete genome
```

```
              Length = 255408

 Score = 64.0 bits (153), Expect = 7e-10
 Identities = 54/170 (31%), Positives = 78/170 (45%), Gaps = 45/170 (26%)
 Frame = -3

Query: 9      PDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQP---QGHKSCGHCRGC 65
              P    L  +  A R  HA L     G G +++ + L++ L C+      G  SCG C  C
Sbjct: 252943 PQLRVLKTALGANRLAHAYLFTGPEGSGKESVAFELAKILNCRSSGNLSGEGSCGECESC 252764

Query: 66     QLMQAGTHPD-------------------------------YYTLAPEKGKN-------- 86
              +      HP+                               Y  L  EK KN
Sbjct: 252763 RQTDLLMHPNIEYLFPVEAALLETIDPSKKENKKLTEARERYEALLDEKRKNPFFTPAME 252584

Query: 87     -TLGV--DAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFL 143
               ++G+  + V  + +K +   R GG KV  +                  EEPPA   F L
Sbjct: 252583 RSMGILTEQVVMLQQKASLAPRDGGKKVFIISQAERLHPTAANKLLKLLEEPPAHVVFIL 252404

Query: 144    ATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSR 178
               +  PE +L T+RSRC+L   A P       W++R
Sbjct: 252403 VSSRPESVLPTIRSRCQLLNFARPRPAEIEAWIAR 252299


 gnl|TIGR|gef_6277 Enterococcus faecalis unfinished fragment of complete genome
          Length = 9336

 Score = 62.8 bits (150), Expect = 2e-09
 Identities = 37/153 (24%), Positives = 66/153 (42%), Gaps = 1/153 (0%)
 Frame = -1

Query: 11   FEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQA 70
            +++L  S++ GR  HA L +   G G       +++++ C      + C  C  C  +
Sbjct: 8865 YKQLQKSFEHGRLAHAYLFEGDTGTGKQEFGLWMAKHVFCTNLVNQQPCNECHNCVRINE 8686

Query: 71   GTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXX 130
              HPD   +AP+    T+ V+ +RE+  + ++       KV  +
Sbjct: 8685 NEHPDVLRIAPD--GQTIKVNQIRELKAEFSKSGVETAKKVFLIQEADKMSTGAANSLLK 8512

Query: 131  XXEEPPAETWFFLATREPERLLATLRSRCR-LHY 163
              EEP  +    L T    R+L T++SRC+ LH+
Sbjct: 8511 FLEEPEGQILAILETTSLSRILPTIQSRCQTLHF 8410


 gnl|Sanger|Y.pesits_Contig790 Yersinia pestis unfinished fragment of complete genome
          Length = 98765

 Score = 62.8 bits (150), Expect = 2e-09
 Identities = 40/144 (27%), Positives = 60/144 (40%)
 Frame = -3

Query: 21    GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
             GR HHA L     G+G  ++   L++ L C+       CG C  CQ ++ G   D   +
Sbjct: 63444 GRIHHAYLFSGTRGVGKTSIARLLAKGLNCETGITATPCGTCANCQEIEQGRFVDLIEI- 63268

Query: 81    PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETW 140
                  +   V+  RE+ + +      G  KV  +                  EEPPA
Sbjct: 63267 --DAASRTKVEDTRELLDNVQYAPARGRFKVYLIDEVHMLSRHSFNALLKTLEEPPAHVK 63094

Query: 141   FFLATREPERLLATLRSRCRLHYL 164
             F LAT +P++L  T+ SRC   +L
Sbjct: 63093 FLLATTDPQKLPVTILSRCLQFHL 63022
```

```
gnl|TIGR|C.trachomatis_ct_97 Chlamydia trachomatis MOPN unfinished fragment of complete
          Length = 4554

 Score = 62.5 bits (149), Expect = 2e-09
 Identities = 41/161 (25%), Positives = 67/161 (41%), Gaps = 1/161 (0%)
 Frame = -2

Query: 17   SYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQP-QGHKSCGHCRGCQLMQAGTHPD 75
            S +  R  HA +    + G G   L    ++ L CQ P +  + C  C  C+ +  GT  D
Sbjct: 1487 SLRLNRSAHAYIFSGIRGTGKTTLARVFAKALNCQSPTENQEPCNQCAICKEISLGTSMD 1308

Query: 76   YYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEP 135
              +     G +  G++ +R++ E +           K+   +                EEP
Sbjct: 1307 VMEI---DGASHRGIEDIRQINETVLXVPSKSRYKIYIIDEVHMLTKEAFNSLLKTLEEP 1137

Query: 136  PAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLS 177
            PA    FFLAT E   ++   T+ SRC+    L    PE+  +   L+
Sbjct: 1136 PAHVKFFLATTEIAKIPNTISSRCQKMLLKRIPEETIIDKLT 1011


 gnl|PAGP|Paeruginosa_Contig53 Pseudomonas aeruginosa unfinished fragment of complete ger
            Length = 1300758

 Score = 62.1 bits (148), Expect = 3e-09
 Identities = 69/268 (25%), Positives = 103/268 (37%), Gaps = 12/268 (4%)
 Frame = +2

Query: 14     LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
              L+ +      R HHA L       G+G     +     L++ L C+          CG C   C+ +  G
Sbjct: 943820 LINALDNQRLHHAYLFTGTRGVGKTTIARILAKCLNCETGVSSTPCGECSVCREIDEGRF 943999

Query: 74     PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXE 133
               D   L       +    V+   RE+ + +       G  KV   +                  E
Sbjct: 944000 VD---LIEVDAASRTKVEDTRELLDNVQYSPTRGRYKVYLIDEVHMLSSHSFNALLKTLE 944170

Query: 134    EPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWL-----SREVTMSQDXXX 188
              EPP      F LAT +P++L   T+ SRC    L    P +  V  L      +  V    D
Sbjct: 944171 EPPPHVKFLLATTDPQKLPVTILSRCLQFSLKNMPPERVVEHLTHVLGAENVPFEDDALW 944350

Query: 189    XXXXXXXXXXXXXXXXFQGDNWQARETLCQALAY---SVPSGDWYSLLAALNHEQA---- 241
                           G     A      QA+A+     V + D   ++L   L+H Q
Sbjct: 944351 LLGRAA-----------DGSMRDAMSLTDQAIAFGEGKVLAADVRAMLGTLDHGQVYGVL 944497

Query: 242    PARLHWLATLLMDALKRHHGAAQVTNVDVPGLVAELANHL 281
               A L   A   L++A++  H  A Q   D  G++AE+  N L
Sbjct: 944498 QALLEGDARALLEAVR--HLAEQ--GPDWGGVLAEILNVL 944605


 gnl|TIGR|N.meningitidis_GNMCF18R Neisseria meningitidis MC58 unfinished fragment of comp
          Length = 639

 Score = 60.9 bits (145), Expect = 6e-09
 Identities = 39/125 (31%), Positives = 52/125 (41%)
 Frame = -1

Query: 21  GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
            GR HHA L+     G+G    +    L++ L C+  Q   + CG C  C   + AG + D    L
Sbjct: 369 GRLHHAYLLTGTRGVGKTTIARILAKSLNCENAQHGEPCGVCESCTQIDAGRYVD--LLE 196
```

```
Query: 81   PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETW 140
              +   NT G+D +REV E         G  KV   +                      EEPP
Sbjct: 195  IDAASNT-GIDNIREVLENAQYAPTAGKYKVYIIDEVHMLSKSAFNAMLKTLEEPPEHVK 19

Query: 141  FFLAT 145
             F LAT
Sbjct: 18   FILAT 4
```

gnl|TIGR|T.maritima_tm_26 Thermotoga maritima unfinished fragment of complete genome
Length = 18920

Score = 60.9 bits (145), Expect = 6e-09
Identities = 37/157 (23%), Positives = 63/157 (39%)
Frame = -2

```
Query: 14    LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
              ++ + Q     H  +     G G    L    L++ L C+   +G + C   CR C+ +  GT
Sbjct: 5536  IIGAIQKNSVAHGYIFAGPRGTGKTTLARILAKSLNCENRKGVEPCNSCRACREIDEGTF 5357

Query: 74    PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXE 133
               D   L      +    G+D +R + + +        G  KV   +                   E
Sbjct: 5356  MDVIELDAASNR---GIDEIRRIRDAVGYRPMEGKYKVYIIDEVHMLTKEAFNALLKTLE 5186

Query: 134   EPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQ 170
              EPP+     F LAT    E++   T+ SRC++        P++
Sbjct: 5185  EPPSHVVFVLATTNLEKVPPTIISRCQVFEFRNIPDE 5075
```

gnl|TIGR|P.gingivalis_1194 Porphyromonas gingivalis W83 unfinished fragment of complete
Length = 418115

Score = 59.7 bits (142), Expect = 1e-08
Identities = 79/303 (26%), Positives = 126/303 (41%), Gaps = 53/303 (17%)
Frame = +2

```
Query: 25      HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP--- 81
                HA L      G G    L  A +RYL CQ P     +CGHC   C      A   HPD + + P
Sbjct: 102800  HAQLFAGEEGGGAFPLALAYARYLNCQMPTDTDACGHCPSCVKYDALAHPDLFFVYPVVN 102979

Query: 82      ------------EKGKNTLGVDA-----------------------VREVTEKLNEHAR 105
                              + +  LG ++                          V +KL+
Sbjct: 102980  ASSSPAPSDDYIRQWREMLGSESYFTPADWLEYIKAGNSQPIIYSKEAEAVEQKLSFRIY 103159

Query: 106     LGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLA 165
                  +VV  +                           EEPP   T FF+  +  EP+++L T+RSR +L   +
Sbjct: 103160  EASYRVVMIWQPERMNEAMANKLLKLIEEPPEHTLFFMISSEPDKVLGTIRSRTQLINVR 103339

Query: 166     GPPEQYAVTWLSREVTMSQDXXXXXXXXXXXXXXXXXXFQGDNWQARE--TLCQALAYS 223
                  E   V  LSR      +                               ++G+ W  R+     L    +   S
Sbjct: 103340  LLHEIEIVEALSRNNQGNTADIIRIAHLAEGNYRRAMDLYRGE-WADRDNFVLMGRMMGS 103516

Query: 224     VPSGDWYSL------LAALNHEQAPARLHWLATLLMDALKRHHGAAQVT--NVDVPGLVA 275
                +  GD     +          LAAL           L +    L  +        G A++      + +       V
Sbjct: 103517  IIKGDPSKMRPVADELAALGRVSQIGFLTYCLRLFRELYISRVGVAKLNYLSPEEESFVD 103696

Query: 276     ELANHLSPSRLQAILGDV----CHIREQLMSVTGINRELLITDLLLRIEHYLQPGV 327
                 L+   ++    ++  ++  +V       HIR+               N   ++   DLLLR+     L  P  +
Sbjct: 103697  MLSGGITGQNIRPVMEEVELAIRHIRQ------NGNGRMIFFDLLLRLTAALAPAL 103846
```

```
gnl|Sanger|S.typhi_Contig376 Salmonella typhi unfinished fragment of complete genome
          Length = 157214

 Score = 59.3 bits (141), Expect = 2e-08
 Identities = 38/144 (26%), Positives = 60/144 (41%)
 Frame = +1

Query: 21    GRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLA 80
              GR HHA L     G+G  ++    L++ L C+        CG C   C+ ++ G    D    +
Sbjct: 13384 GRIHHAYLFSGTRGVGKTSIARLLAKGLNCETGITATPCGVCDNCREIEQGRFVDLIEI- 13560

Query: 81    PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETW 140
                  +   V+  R++ + +        G  KV   +                      EEPPA
Sbjct: 13561 --DAASRTKVEDTRDLLDNVQYAPARGRFKVYLIDEVHMLSRHSFNALLKTLEEPPAHVK 13734

Query: 141   FFLATREPERLLATLRSRCRLHYL 164
              F LAT +P++L   T+ SRC    +L
Sbjct: 13735 FLLATTDPQKLPVTILSRCLQFHL 13806


gnl|OUACGT|Spyogenes_Contig243 Streptococcus pyogenes unfinished fragment of complete ge
          Length = 22344

 Score = 59.3 bits (141), Expect = 2e-08
 Identities = 34/140 (24%), Positives = 60/140 (42%)
 Frame = +1

Query: 22    RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
              R +HA L          ++ +    L++ + C+Q +     CGHCR CQL++ G    D    L P
Sbjct: 17944 RLNHAYLFSG--DFANEEMALFLAKVIFCEQKKDQTPCGHCRSCQLIEQGDFADVTVLEP 18117

Query: 82    EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWF 141
                   +  D V+E+      ++         +V  +                     EEP   E +
Sbjct: 18118 T--GQVIKTDVVKEMMANFSQTGYENKRQVFIIKDCDKMHINAANSLLKYIEEPQGEAYI 18291

Query: 142   FLATREPERLLATLRSRCRL 161
              FL T +   ++L T++SR ++
Sbjct: 18292 FLLTNDDNKVLPTIKSRTQV 18351


gnl|TIGR|S.putrefaciens_gsp_387 Shewanella putrefaciens unfinished fragment of complete
          Length = 3834

 Score = 58.9 bits (140), Expect = 2e-08
 Identities = 44/163 (26%), Positives = 61/163 (36%)
 Frame = +1

Query: 22  RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
            R HHA L     G+G  +L     ++ L C+        CG C   C  +  G   D   L
Sbjct: 562 RLHHAYLFTGTRGVGKTSLARLFAKGLNCETGVTASPCGVCGSCVEIAQGRFVD---LIE 732

Query: 82  EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWF 141
                +   VD   RE+ + +        G  KV   +                      EEPP    F
Sbjct: 733 VDAASRTKVDDTRELLDNVQYRPTRGRFKVYLIDEVHMLSRSSFNALLKTLEEPPEHVKF 912

Query: 142 FLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSREVTMSQ 184
             LAT +P++L   T+ SRC     L     +Q     T L     +T   Q
Sbjct: 913 LLATTDPQKLPVTVLSRCLQFNLKSLTQQEIGTQLQHILTQEQ 1041


gnl|TIGR|M.avium_5593 Mycobacterium avium unfinished fragment of complete genome
```

```
             Length = 21394

 Score = 58.2 bits (138), Expect = 4e-08
 Identities = 46/152 (30%), Positives = 60/152 (39%)
 Frame = +2

Query: 12    EKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAG 71
             E L  + +AGR +HA L      G G  +     L+R L C Q       CG C  C L  A
Sbjct: 9860  EPLSIALEAGRINHAYLFSGPRGCGKTSSARILARSLNCVQGPTATPCGVCDSC-LALAP 10036

Query: 72    THPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXX 131
               P     +       +  GVD  RE+ ++              +V  V
Sbjct: 10037 NAPGSIDVVELDAASHGGVDDTRELRDRAFYAPAQSRYRVFIVDEAHMVTTAGFNALLKI 10216

Query: 132   XEEPPAETWFFLATREPERLLATLRSRCRLHY 163
              EEPP     F  AT EPE++L T+RSR    HY
Sbjct: 10217 VEEPPEHLIFIFATTEPEKVLPTIRSRTH-HY 10309


 emb|AL009126|BSUB Bacillus subtilis complete genome
             Length = 4214814

 Score = 58.2 bits (138), Expect = 4e-08
 Identities = 43/154 (27%), Positives = 72/154 (45%), Gaps = 3/154 (1%)
 Frame = +1

Query: 7     LRPDFEKLVA-SYQAGRGHHALLIQALPGMG--DDALIYALSRYLLCQQPQGHKSCGHCR 63
             L+P    KL+   S +   R  HA L +    G G  D AL+ A S + L      G + C  CR
Sbjct: 40693 LQPRVMKLLYNSIEKDRLSHAYLFEGKKGTGKLDAALLLAKSFFCL---EGGAEPCESCR 40863

Query: 64    GCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXX 123
               C+ +++G HPD + + P+      ++      ++ + E+ ++          K+   +
Sbjct: 40864 NCKRIESGNHPDLHLVQPD--GLSIKKAQIQALQEEFSKTGLESHKKLYIISHADQMTAN 41037

Query: 124   XXXXXXXXXEEPPAETWFFLATREPERLLATLRSRCR 160
                          EEP  +T    L T +P+RLL T+ SRC+
Sbjct: 41038 AANSLLKFLEEPNKDTMAVLITEQPQRLLDTIISRCQ 41148


 Score = 49.6 bits (116), Expect = 1e-05
 Identities = 32/136 (23%), Positives = 52/136 (37%)
 Frame = +1

Query: 25    HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
             HA L     G G  +      ++ + C+       + C  C  C+ +  G+  D    +
Sbjct: 26926 HAYLFSGPRGTGKTSAAKIFAKAVNCEHAPVDEPCNECAACKGITNGSISDVIEIDAASN 27105

Query: 85    KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                 GVD +R++ +K+            KV  +                    EEPP     F LA
Sbjct: 27106 N---GVDEIRDIRDKVKFAPSAVTYKVYIIDEVHMLSIGAFNALLKTLEEPPEHCIFILA 27276

Query: 145   TREPERLLATLRSRCR 160
             T EP ++   T+ SRC+
Sbjct: 27277 TTEPHKIPLTIISRCQ 27324


 gnl|UOKNOR|S.mutans_Contig840 Streptococcus mutans unfinished fragment of complete genon
           Length = 5373

 Score = 58.2 bits (138), Expect = 4e-08
 Identities = 36/153 (23%), Positives = 66/153 (42%)
```

```
 Frame = -1

Query: 11   FEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQA 70
            F++     Q+G+  HA L         + A++ A SR+  C  P     CG CR C+L+
Sbjct: 4308 FQEFQRILQSGKLSHAYLFSGDFASFEMAVLLAQSRF--CDSPIDALPCGQCRSCRLIAE 4135

Query: 71   GTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXX 130
               D   + PE      +     +R++  + +       G ++V  +
Sbjct: 4134 NDFSDVKVIEPE--GQMIKTATIRDLLREFSSSGFEGQSQVFIIRDADKMHTNAANSLLK 3961

Query: 131  XXEEPPAETWFFLATREPERLLATLRSRCRLHY 163
              EEP ++T+   L T++   R+L T++SR ++ Y
Sbjct: 3960 FIEEPQSDTYMILLTQDESRILPTIKSRTQIFY 3862


 gb|AE001273|AE001273 Chlamydia trachomatis complete genome
              Length = 1042519

 Score = 57.8 bits (137), Expect = 5e-08
 Identities = 37/144 (25%), Positives = 59/144 (40%), Gaps = 1/144 (0%)
 Frame = +2

Query: 17     SYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQP-QGHKSCGHCRGCQLMQAGTHPD 75
              S +  R  HA +   + G G   L     ++ L CQ P Q  + C  C  C+ +  GT  D
Sbjct: 381368 SLRLNRAAHAYIFSGIRGTGKTTLARVFAKALNCQNPTQDQEPCNQCAICKEISLGTSMD 381547

Query: 76     YYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEP 135
                +     G +  G++ +R++ E +            K+  +                  EEP
Sbjct: 381548 VIEI---DGASHRGIEDIRQINETVLFVPSKSRYKIYIIDEVHMLTKEAFNSLLKTLEEP 381718

Query: 136    PAETWFFLATREPERLLATLRSRCR 160
              P     FFLAT E  ++   T+ SRC+
Sbjct: 381719 PVHVKFFLATTEIAKIPNTISSRCQ 381793


 Score = 35.6 bits (80), Expect = 0.25
 Identities = 20/87 (22%), Positives = 34/87 (38%)
 Frame = -1

Query: 73     HPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXX 132
              HPD +   +P+         ++   R + + +  H       K+   +
Sbjct: 209608 HPDMHEYSPQGKGRLHTIETPRAIRKDIWIHPYESPYKIYIIYEADRITLDAISAFLKLL 209429

Query: 133    EEPPAETWFFLATREPERLLATLRSRC 159
              E+PP     F L +  P+RL  T+RSRC
Sbjct: 209428 EDPPQYGMFILVSALPQRLPPTIRSRC 209348


 gnl|TIGR|C.crescentus_gcc_764 Caulobacter crescentus unfinished fragment of complete ger
          Length = 943

 Score = 57.0 bits (135), Expect = 9e-08
 Identities = 43/164 (26%), Positives = 68/164 (41%), Gaps = 8/164 (4%)
 Frame = +3

Query: 15  VASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQ----A 70
            + + + GR HHA L+    G+G    L Y ++R LL  +P    +           ++     A
Sbjct: 366 IDALERGRLHHAWLLTGPEGVGKATLAYRMARRLLGARPDPSQGLLGAAPSDVVSRQVAA 545

Query: 71  GTHPDYYTLA----PEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXX 126
             +HPD    L        K + ++ VD  R++ E        +   +V  +
```

```
Sbjct: 546 RSHPDLMVLERLTDDGKARKSIPVDEARKLPEFFANSPAVSPYRVAIIDAADDLNVNAAN 725

Query: 127 XXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQYAVTWLSR 178
                 EEPPA      L +   P +LL T+RSRCR    +   P     A     + R
Sbjct: 726 AVLKTLEEPPARGVILLISHAPGKLLPTIRSRCRRLAIPAPGVAAAAXMVER 881


 gnl|Sanger|campylo_Cj.seq Campylobacter jejuni NCTC 11168 unfinished fragment of complet
              Length = 1641480

 Score = 56.6 bits (134), Expect = 1e-07
 Identities = 38/134 (28%), Positives = 53/134 (39%)
 Frame = +2

Query: 25      HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
            HA L    L G G   +       SR L+C+Q        CG C+ C       G H D    +
Sbjct: 1089785 HAYLFSGLRGSGKTSSARIFSRALVCEQGPSDTPCGTCKHCLAALEGKHIDIIEMDAASN 1089964

Query: 85      KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
            +      + A+ E T+       AR      K+   +                          EEPP+       F LA
Sbjct: 1089965 RGLEDIQALIEQTKYTPSMARF---KIFIIDEVHMLTPQAANALLKTLEEPPSYVKFILA 1090135

Query: 145     TREPERLLATLRSR 158
            T +P +L AT+ SR
Sbjct: 1090136 TTDPLKLPATVLSR 1090177


 gnl|OUACGT|A.actin_Contig753 Actinobacillus actinomycetemcomitans unfinished fragment of
          genome
          Length = 7256

 Score = 56.2 bits (133), Expect = 2e-07
 Identities = 38/151 (25%), Positives = 58/151 (38%)
 Frame = -1

Query: 14   LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
          L     +   R HHA L       G+G   ++       ++ L C        + CG C   C   ++ G
Sbjct: 4589 LANGLRENRLHHAYLFSGTRGVGKTSIARLFAKGLNCVSGVTAEPCGVCEHCNAIEKGNF 4410

Query: 74   PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXE 133
           D    +           +    V+    RE+ + +       LG    KV    +                           E
Sbjct: 4409 IDLIEI---DAASRTKVEDTRELLDNVQYKPVLGRYKVYLIDEVHMLSRHSFNALLKTLE 4239

Query: 134  EPPAETWFFLATREPERLLATLRSRCRLHYL 164
          EPP       F LAT +P +L   T+ SRC     +L
Sbjct: 4238 EPPEYVKFLLATTDPHKLPVTILSRCMQFHL 4146


 gnl|TIGR|gmt3732 Mycobacterium tuberculosis unfinished fragment of complete genome
            Length = 466170

 Score = 55.8 bits (132), Expect = 2e-07
 Identities = 44/150 (29%), Positives = 58/150 (38%)
 Frame = +3

Query: 14      LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
            L  +   AGR +HA L       G G   +       L+R L C Q           CG C   C   + A
Sbjct: 394476 LSVALDAGRINHAYLFSGPRGCGKTSSARILARSLNCAQGPTANPCGVCESCVSL-APNA 394652

Query: 74      PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXE 133
            P      +           +   GVD   RE+ ++                +V   V                               E
```

```
Sbjct: 394653 PGSIDVVELDAASHGGVDDTRELRDRAFYAPVQSRYRVFIVDEAHMVTTAGFNALLKIVE 394832

Query: 134     EPPAETWFFLATREPERLLATLRSRCRLHY 163
               EPP     F  AT EPE++L T+RSR    HY
Sbjct: 394833 EPPEHLIFIFATTEPEKVLPTIRSRTH-HY 394919


 gnl|GTC|C.aceto_AE001437 Clostridium acetobutylicum, WORKING DRAFT SEQUENCE, 1 ordered p
              Length = 3943874

 Score = 55.4 bits (131), Expect = 3e-07
 Identities = 36/136 (26%), Positives = 53/136 (38%)
 Frame = +3

Query: 25      HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
                HA L+     G G        LS+ + C   PQ   + C  C  C+ + AG    D    L
Sbjct: 3734367 HAYLMCGTRGTGKTTTAKILSKAVNCLNPQDGEPCNECEMCKKINAGIAIDVTELDAASN 3734546

Query: 85      KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                 +    VD +R + + +            KV   +                   EEPP     F LA
Sbjct: 3734547 NS---VDDIRNIIDDVQYPPHESKFKVYIIDEVHMLSQGAVNAFLKTLEEPPQNVVFILA 3734717

Query: 145     TREPERLLATLRSRCR 160
                T +P++L   T+ SRC+
Sbjct: 3734718 TTDPQKLPVTILSRCQ 3734765


 Score = 44.1 bits (102), Expect = 6e-04
 Identities = 23/80 (28%), Positives = 39/80 (48%)
 Frame = +3

Query: 85    KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
              K ++ VD VR++ E++N+      G   K++ V                         EEPP     +   L
Sbjct: 14097 KKSISVDQVRKIIEEVNKKPYEGNNKLIVVHDMDYMTIQGQNAFLKTIEEPPLGVYIILL 14276

Query: 145   TREPERLLATLRSRCRLHYL 164
               +     R+L T+RSRC+++ L
Sbjct: 14277 CQSQGRVLDTVRSRCQIYKL 14336


 gnl|TIGR|S.pneumoniae_sp_36 Streptococcus pneumoniae unfinished fragment of complete ger
              Length = 43015

 Score = 54.7 bits (129), Expect = 4e-07
 Identities = 35/165 (21%), Positives = 66/165 (39%)
 Frame = +1

Query: 6     WLRPDFEKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGC 65
              W      F++ V     +   + +HA L                 +     L++ L C      G     C   CR C
Sbjct: 23515 WQPAQFDRFVRILEQDQLNHAYLFSGF--FESLEMAQFLAKSLFCTDKVGVLPCEKCRSC 23688

Query: 66    QLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXX 125
              +L++ G    PD     + P           +   + +RE+   + ++           +V   +
Sbjct: 23689 KLIEQGEFPDVTLIKPV--NQVIKTERIRELVGQFSQAGIESQQQVFIIEQADKMHPNAA 23862

Query: 126   XXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHYLAGPPEQ 170
                     EEP +E + F   T + E++L T+RSR ++ +        E+
Sbjct: 23863 NSLLKVIEEPQSEVYIFFLTSDEEKMLPTIRSRTQIFHFKKQEEK 23997


 gnl|TIGR|V.cholerae_asm864 Vibrio cholerae unfinished fragment of complete genome
```

```
               Length = 23778

 Score = 54.3 bits (128), Expect = 6e-07
 Identities = 37/143 (25%), Positives = 55/143 (37%)
 Frame = -3

Query: 22    RGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAP 81
             R HHA L     G+G    +   ++ L C+       CG C  CQ +  G   D   +
Sbjct: 14509 RLHHAYLFSGTRGVGKTTIGRLFAKGLNCETGITATPCGQCATCQEIDQGRFVDLLEI-- 14336

Query: 82    EKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWF 141
                 +   V+  RE+ + +      G  KV  +                  EEPP    F
Sbjct: 14335 -DAASRTKVEDTRELLDNVQYKPARGRFKVYLIDEVHMLSRHSFNALLKTLEEPPEYVKF 14159

Query: 142   FLATREPERLLATLRSRCRLHYL 164
              LAT +P++L  T+ SRC   +L
Sbjct: 14158 LLATTDPQKLPVTILSRCLQFHL 14090


 gnl|TIGR|P.gingivalis_1209 Porphyromonas gingivalis W83 unfinished fragment of complete
               Length = 276255

 Score = 53.9 bits (127), Expect = 8e-07
 Identities = 36/137 (26%), Positives = 59/137 (42%), Gaps = 2/137 (1%)
 Frame = +2

Query: 25    HALLIQALPGMGDDALIYALSRYLLCQQ--PQGHKSCGHCRGCQLMQAGTHPDYYTLAPE 82
             HA L     G+G  +     +R + C +  P G ++CG C  C+         + Y L
Sbjct: 15524 HAYLFCGPRGVGKTSCARIFARAINCLERLPDG-EACGRCESCKAFDEQRSMNIYELDAA 15700

Query: 83    KGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFF 142
                +   VD +R + E+ N   ++G  K+  +                  EEPP+   F
Sbjct: 15701 SNNS---VDDIRLLIEQANVPPQIGKYKIYIIDEVHMLSQQAFNAFLKTLEEPPSYVIFI 15871

Query: 143   LATREPERLLATLRSRCRL 161
             LAT E  ++L T+ SRC++
Sbjct: 15872 LATTEKHKILPTILSRCQI 15928


 gnl|Sanger_1765|mbovis_Contig454.1 Mycobacterium bovis unfinished fragment of complete g
              Length = 1934

 Score = 53.5 bits (126), Expect = 1e-06
 Identities = 42/150 (28%), Positives = 57/150 (38%)
 Frame = -3

Query: 14   LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
            L  +  AGR +HA L     G G  +    L+R L C Q      CG C  C +
Sbjct: 1185 LSVALDAGRINHAYLFSGPRGCGKTSSARILARSLNCAQGPTANPCGVCESCVSLAPNAL 1006

Query: 74   PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXE 133
                 +  +   +  GVD  RE+ ++          +V  V                  E
Sbjct: 1005 GSIDVVELDAASHG-GVDDTRELRDRAFYAPVQSRYRVFIVDEAHMVTTAGFNALLKIVE 829

Query: 134  EPPAETWFFLATREPERLLATLRSRCRLHY 163
            EPP    F  AT EPE++L T+RSR   HY
Sbjct: 828  EPPEHLIFIFATTEPEKVLPTIRSRTH-HY 742


 gb|AE000520|AE000520 Treponema pallidum complete genome
                 Length = 1138011
```

```
 Score = 53.1 bits (125), Expect = 1e-06
 Identities = 38/147 (25%), Positives = 60/147 (39%)
 Frame = -3

Query: 14      LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
                L  S +  +   A L     G G  +    L++ L C Q +  + CG C  C+ +  GT+
Sbjct: 1094869 LQKSLEENKVSPAYLFSGPHGCGKTSCARILAKALNCVQREASEPCGECPSCREIATGTN 1094690

Query: 74      PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXE 133
                 +   +    G +  GV  VR++ E++         KV  +                  E
Sbjct: 1094689 LNVIEI---DGASHTGVGDVRQIKEEILFPPHGTRYKVFIIDEVHMLSNSAFNALLKTIE 1094519

Query: 134     EPPAETWFFLATREPERLLATLRSRCR 160
                EPP    F  AT E  R+ AT++SRC+
Sbjct: 1094518 EPPPYVVFIFATTEVHRIPATVKSRCQ 1094438


 gb|AE000511|HPYL Helicobacter pylori 26695 complete genome
           Length = 1667867

 Score = 51.5 bits (121), Expect = 4e-06
 Identities = 38/152 (25%), Positives = 58/152 (38%)
 Frame = -1

Query: 25     HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
              +A L   L G G  +     +R L+C++      C  C  CQ      H D   +    G
Sbjct: 772097 NAYLFSGLRGSGKTSSSRIFARALMCEEGPKAVPCDTCIQCQSALNNHHIDIIEM---DG 771927

Query: 85     KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
               +  G+D VR + E+       G  K+  +                  EEPP+   F LA
Sbjct: 771926 ASNRGIDDVRNLIEQTRYKPSFGRYKIFIIDEVHMFTTEAFNALLKTLEEPPSHVKFLLA 771747

Query: 145    TREPERLLATLRSRCRLHYLAGPPEQYAVTWL 176
              T +  +L AT+ SR +       PE   ++ L
Sbjct: 771746 TTDALKLPATILSRTQHFRFKKIPENSVISHL 771651


 gnl|TIGR|S.aureus_2202 Staphylococcus aureus COL unfinished fragment of complete genome
          Length = 30502

 Score = 51.2 bits (120), Expect = 5e-06
 Identities = 32/136 (23%), Positives = 52/136 (37%)
 Frame = -3

Query: 25   HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
            HA +     G G  ++    ++ + C      + C  C  C+ +  GT+ D   +
Sbjct: 5951 HAYIFSGPRGTGKTSIAKVFAKAINCLNSTDGEPCNECHICKGITQGTNSDVIEIDAASN 5772

Query: 85   KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                GVD +R + +K+         KV  +                  EEPPA   F LA
Sbjct: 5771 N---GVDEIRNIRDKVKYAPSESKYKVYIIDEVHMLTTGAFNALLKTLEEPPAHAIFILA 5601

Query: 145  TREPERLLATLRSRCR 160
            T EP ++  T+ SR +
Sbjct: 5600 TTEPHKIPPTIISRAQ 5553


 gnl|OUACGT|S.aureus_Contig1164 Staphylococcus aureus unfinished fragment of complete ger
         Length = 1224
```

```
 Score = 51.2 bits (120), Expect = 5e-06
 Identities = 32/136 (23%), Positives = 52/136 (37%)
 Frame = +2

Query: 25   HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
            HA +      G G  ++     ++ + C        + C  C  C+ +  GT+ D    +
Sbjct: 740  HAYIFSGPRGTGKTSIAKVFAKAINCLNSTDGEPCNECHICKGITQGTNSDVIEIDAASN 919

Query: 85   KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                GVD +R + +K+             KV  +                  EEPPA    F LA
Sbjct: 920  N---GVDEIRNIRDKVKYAPSESKYKVYIIDEVHMLTTGAFNALLKTLEEPPAHAIFILA 1090

Query: 145  TREPERLLATLRSRCR 160
            T EP ++   T+ SR +
Sbjct: 1091TTEPHKIPPTIISRAQ 1138


 gb|AE001439|AE001439 Helicobacter pylori, strain J99 complete genome
            Length = 1643831

 Score = 50.0 bits (117), Expect = 1e-05
 Identities = 38/152 (25%), Positives = 57/152 (37%)
 Frame = -3

Query: 25      HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
               +A L     L G G  +       +R L+C+            C  C  CQ        H D    +     G
Sbjct: 734547 NAYLFSGLRGSGKTSSSRIFARALMCKTGPKAVPCDTCIQCQSALNNHHIDIIEM---DG 734377

Query: 85      KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                +   G+D VR + E+          G  K+   +                   EEPP+    F LA
Sbjct: 734376 ASNRGIDDVRNLIEQTRYKPSFGRYKIFIIDEVHMFTTEAFNALLKTLEEPPSHVKFLLA 734197

Query: 145     TREPERLLATLRSRCRLHYLAGPPEQYAVTWL 176
               T +  +L AT+ SR +          PE    ++ L
Sbjct: 734196 TTDALKLPATILSRTQHFRFKKIPENSVISHL 734101


 gnl|TIGR|N.meningitidis_GNMAB03R Neisseria meningitidis MC58 unfinished fragment of comp
          Length = 435

 Score = 49.6 bits (116), Expect = 1e-05
 Identities = 32/116 (27%), Positives = 50/116 (42%)
 Frame = +1

Query: 44   LSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEH 103
            L++ L C+  Q   + CG C+ C   + AG + D    +   +   NT G+D +REV E
Sbjct: 58   LAKSLNCENAQHGEPCGVCKSCTQIDAGRYVDLLEI--DAASNT-GIDNIREVLENAQYA 228

Query: 104  ARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRC 159
                G  KV  +                    +  P   + F LAT +P ++   T+ SRC
Sbjct: 229  PTAGKYKVYIIDEGICFPKARSTLCSKRWKSRPNTSKFILATTDPHKVPVTVLSRC 396


 gnl|TIGR|S.pneumoniae_sp_68 Streptococcus pneumoniae unfinished fragment of complete ger
           Length = 21744

 Score = 49.2 bits (115), Expect = 2e-05
 Identities = 34/134 (25%), Positives = 51/134 (37%)
 Frame = -3

Query: 25     HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
```

```
                HA L     G G   ++     ++ + C    G + C +C  CQ +  G+  D    +
Sbjct: 17440 HAYLFSGPRGTGKTSVAKIFAKAMNCPNQVGGEPCNNCYICQAVTDGSLEDVIEMDAASN 17261

Query: 85    KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                 GVD +RE+ +K        L    KV  +                     EEP      F LA
Sbjct: 17260 N---GVDEIREIRDKSTYAPSLARYKVYIIDEVHMLSTGAFNALLKTLEEPTQNVVFILA 17090

Query: 145   TREPERLLATLRSR 158
              T E  ++ AT+ SR
Sbjct: 17089 TTELHKIPATILSR 17048


 gnl|TIGR|C.tepidum_gct_35 Chlorobium tepidum unfinished fragment of complete genome
             Length = 33899

 Score = 48.0 bits (112), Expect = 4e-05
 Identities = 37/144 (25%), Positives = 58/144 (39%), Gaps = 8/144 (5%)
 Frame = +1

Query: 17    SYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQ----PQGHKS----CGHCRGCQLM 68
              S + GR  H  +   L G+G        ++ + CQ+     PQ  K      CG C  C+
Sbjct: 29680 SLRMGRVGHGYIFSGLRGVGKTTAARVFAKAVNCQRMIDDPQYLKEVTEPCGVCESCRDF 29859

Query: 69    QAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXX 128
               AG        ++      +   VD +R + E +     + G  +V  +
Sbjct: 29860 DAGAS---LNISEFDAASNNSVDDIRLLRENVRYGPQKGRYRVYIIDEVHMLSTAAFNAF 30030

Query: 129   XXXXEEPPAETWFFLATREPERLLATLRSRCR 160
                  EEPP     F  AT E   ++ AT+ SRC+
Sbjct: 30031 LKTLEEPPPHAIFIFATTELHKIPATIASRCQ 30126


 gnl|TIGR|t_ferrooxidans_64 Thiobacillus ferrooxidans unfinished fragment of complete ger
            Length = 4609

 Score = 48.0 bits (112), Expect = 4e-05
 Identities = 29/115 (25%), Positives = 45/115 (38%)
 Frame = -3

Query: 45   SRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHA 104
             ++ L C++        CG C  C+ + AG    D  L       +   VD  R++ + +
Sbjct: 4607 AKCLNCERGVSSNPCGECSACRSIAAGNFVD---LLEVDAASRTRVDETRDLLDNVQYAP 4437

Query: 105  RLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLATREPERLLATLRSRC 159
              G  K   +                      EEPP     F LAT +P++L   T+ SRC
Sbjct: 4436 TAGRYKAYLIDEVHMLSAHSFNALLKTLEEPPEHVKFLLATTDPQKLPITVLSRC 4272


 gnl|OUACGT|Ngon_Contig196 Neisseria gonorrhoeae unfinished fragment of complete genome
             Length = 23501

 Score = 47.3 bits (110), Expect = 7e-05
 Identities = 24/87 (27%), Positives = 41/87 (46%)
 Frame = -1

Query: 90    VDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLATREPE 149
              +DAVRE+ + +    +  GG +V+ +                     EEPP +   F L +    +
Sbjct: 23495 IDAVREIIDNVYLTSVRGGLRVILIHPAESMNVQAANSLLKVLEEPPPQVVFLLVSHAAD 23316

Query: 150   RLLATLRSRCRLHYLAGPPEQYAVTWL 176
              ++L T++SRCR   L  P    A+ +L
```

```
Sbjct: 23315 KVLPTIKSRCRKMVLPAPSHGEALAYL 23235


 gnl|TIGR|D.radiodurans_8813 Deinococcus radiodurans unfinished fragment of complete genc
             Length = 83236

 Score = 47.3 bits (110), Expect = 7e-05
 Identities = 40/136 (29%), Positives = 59/136 (42%), Gaps = 20/136 (14%)
 Frame = -3

Query: 23    GHHALLIQALPGMGDDALIYALSRYLLCQQPQGH--KSCGHCRGCQLMQAGTHPDYYTLA 80
             G +ALL+     +G    L YA++      C   P+G    ++CG C   C+ +QAG HPD    L
Sbjct: 54530 GGNALLLSGPARVGKLDLAYAIAAQHNCSGPRGMYGEACGQCPSCRALQAGAHPDVLRLE 54351

Query: 81    PEKGKNT--------LGVDAVREVTEKLNEHAR----------LGGAKVVWVXXXXXXXX 122
             P     +T          + + AV E   +     E+                  +VV V
Sbjct: 54350 PRATTSTGKAARKRIIPIGAVLESRDTGREYETHVYEFLEVRPTFERRVVIVAGAEYLNP 54171

Query: 123   XXXXXXXXXXEEPPAETWFFLATREPERLLATLRSR 158
                       EEPP      F       +      +L  T++  SR
Sbjct: 54170 QAANALLKLVEEPPHRALFLFLAEDLRSVLPTIVSR 54063


 gb|AE000783|AE000783 Borrelia burgdorferi complete genome
             Length = 910724

 Score = 47.3 bits (110), Expect = 7e-05
 Identities = 32/149 (21%), Positives = 59/149 (39%)
 Frame = +2

Query: 12     EKLVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAG 71
              E L   S +   +   +A +       G+G    +      A +R L C+          CG C    C+ ++
Sbjct: 482678 ETLKHSIEKNKIANAYIFSGPRGVGKTSSARAFARCLNCRNGPTVMPCGECSNCKSIEND 482857

Query: 72     THPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXX 131
              +   D      +       G +      V    +R++ E++           +       ++     +
Sbjct: 482858 SSLDVVEI---DGASNTSVQDIRQIKEEIMFPPAISKYRIYIIDEVHMLSNSAFNALLKT 483028

Query: 132    XEEPPAETWFFLATREPERLLATLRSRCR 160
               EEPP       F    AT  E   +L    T++SRC+
Sbjct: 483029 IEEPPNYIVFIFATTESHKLPETIKSRCQ 483115


 gnl|TIGR|t_ferrooxidans_1967 Thiobacillus ferrooxidans unfinished fragment of complete g
          Length = 563

 Score = 45.7 bits (106), Expect = 2e-04
 Identities = 23/57 (40%), Positives = 30/57 (52%), Gaps = 1/57 (1%)
 Frame = -3

Query: 44  LSRYLLCQQPQGHK-SCGHCRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKL 100
            L +   LC   P       CG CR  C+L+   G HPD     + PE GK    + ++AVR    E L
Sbjct: 558 LQQVALCFAPTAQGLPCGTCRSCRLLAEGNHPDLLMITPETGKR-IAIEAVRHANEFL 388


 gnl|TIGR|gef_6250 Enterococcus faecalis unfinished fragment of complete genome
           Length = 24587

 Score = 45.3 bits (105), Expect = 3e-04
 Identities = 31/134 (23%), Positives = 48/134 (35%)
 Frame = -2
```

```
Query: 25    HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
             HA L     G G  +     ++ + C+  Q  + C  C  C  + G   D   +
Sbjct: 5419 HAYLFTGPRGTGKTSAAKIFAKAINCKHSQDGEPCNVCETCVAITEGRLNDVIEIDAASN 5240

Query: 85   KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                GV+  +R++  +K            KV   +                    EEPP    F LA
Sbjct: 5239 N---GVEEIRDIRDKAKYAPTQAEYKVYIIDEVHMLSTGAFNALLKTLEEPPQNVIFILA 5069

Query: 145  TREPERLLATLRSR 158
             T EP ++   T+ SR
Sbjct: 5068 TTEPHKIPLTIISR 5027
```

emb|AJ235269|RPXX0 Rickettsia prowazekii strain Madrid E, complete genome
Length = 1111523

Score = 42.2 bits (97), Expect = 0.003
Identities = 29/137 (21%), Positives = 52/137 (37%), Gaps = 4/137 (2%)
Frame = +2

```
Query: 28       LIQALPGMGDDALIYALSRYLLCQ----QPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEK 83
                L+   + G+G        +++ + C       +     K+C  C  C          HPD     +
Sbjct: 1091072 LLTGIRGIGKTTSARIIAKAVNCSALITENTAIKTCEKCTNCVSFNNHNHPDIIEI---D 1091242

Query: 84       GKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXEEPPAETWFFL 143
                  +     +D +R + E           G  K+   +                        EEPP      F
Sbjct: 1091243 AASKTSIDDIRRIIESAEYKPLQGKHKIFIIDEVHMLSKGAFNALLKTLEEPPPHVIFIF 1091422

Query: 144      ATREPERLLATLRSRCRLHYL 164
                AT E +++  +T+ SRC+ + L
Sbjct: 1091423 ATTEVQKVPSTIISRCQRYDL 1091485
```

gnl|OUACGT|Spyogenes_Contig260 Streptococcus pyogenes unfinished fragment of complete ge
Length = 36214

Score = 41.8 bits (96), Expect = 0.003
Identities = 32/145 (22%), Positives = 53/145 (36%)
Frame = +3

```
Query: 14     LVASYQAGRGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTH 73
              L  + ++G+   HA L      G G  +      ++ + C        + C  C  C+ +  G+
Sbjct: 33432 LKQAVESGKISHAYLFSGPRGTGKTSAAKIFAKAMNCPNQVDGEPCNQCDICRDITNGSL 33611

Query: 74     PDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXXE 133
               D  +            GVD +R++ +K               KV  +                  E
Sbjct: 33612 EDVIEIDAASNN---GVDEIRDIRDKSTYAPSRATYKVYIIDEVHMLSTGAFNALLKTLE 33782

Query: 134    EPPAETWFFLATREPERLLATLRSR 158
              EP      F LAT E  ++ AT+ SR
Sbjct: 33783 EPTENVVFILATTELHKIPATILSR 33857
```

gnl|OUACGT|Ngon_Contig166 Neisseria gonorrhoeae unfinished fragment of complete genome
Length = 9825

Score = 41.4 bits (95), Expect = 0.004
Identities = 16/37 (43%), Positives = 25/37 (67%)
Frame = +1

```
Query: 4    YPWLRPDFEKLVASYQAGRGHHALLIQALPGMGDDAL 40
            YPWL P + ++   ++ G GHHA+LI+A  G+G + L
Sbjct: 4321 YPWLMPIYHQIAQTFDEGLGHHAVLIKADAGLGVERL 4431


 gnl|UOKNOR|S.mutans_Contig762 Streptococcus mutans unfinished fragment of complete genon
            Length = 3001

 Score = 40.6 bits (93), Expect = 0.007
 Identities = 31/134 (23%), Positives = 47/134 (34%)
 Frame = -1

Query: 25   HALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRGCQLMQAGTHPDYYTLAPEKG 84
            HA L     G G  +     ++ + C      + C +C  C  +  G+   D   +
Sbjct: 1519 HAYLFSGPRGTGKTSAAKIFAKAMNCPHQADGEPCNNCDICHDITNGSLEDVIEIDAASN 1340

Query: 85   KNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLA 144
                GVD +RE+ +K          KV  +                  EEP     F LA
Sbjct: 1339 N---GVDEIREIRDKSTYAPSRATYKVYIIDEVHMLSTGAFNALLKTLEEPTENVVFILA 1169

Query: 145  TREPERLLATLRSR 158
            T E  ++ AT+ SR
Sbjct: 1168 TTELHKIPATILSR 1127


 gnl|TIGR|C.trachomatis_ct_26 Chlamydia trachomatis MOPN unfinished fragment of complete
           Length = 2341

 Score = 38.7 bits (88), Expect = 0.028
 Identities = 21/87 (24%), Positives = 35/87 (40%)
 Frame = +2

Query: 73  HPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXX 132
           HPD Y  +P+       ++   R + + +  H      K+  +
Sbjct: 905 HPDIYEYSPQGKGRLHTIETPRAIRKNIWIHPYESSYKIYIIYEADRISLDAISAFLKLL 1084

Query: 133 EEPPAETWFFLATREPERLLATLRSRC 159
           E+PP  + F L +  P+RL  T+RSRC
Sbjct: 1085EDPPYYSIFILVSALPQRLPPTIRSRC 1165


 gnl|TIGR|S.aureus_2184 Staphylococcus aureus COL unfinished fragment of complete genome
           Length = 12112

 Score = 38.3 bits (87), Expect = 0.037
 Identities = 35/152 (23%), Positives = 64/152 (42%), Gaps = 6/152 (3%)
 Frame = -3

Query: 12  EKLVASYQAGRGHHALLIQALPGMGDDA-----LIYALSRYLLCQQPQGHKSCGHCRGCQ 66
           ++L  +Y + +  HA L +     GDDA     +    ++ +LCQ             C+
Sbjct: 974 QQLTNAYHSNKLSHAYLFE-----GDDAQTMKQVAINFAKLILCQTDXQ---------CE 837

Query: 67  L-MQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXX 125
              +    HPD+  ++  +  N +  + V ++   +N+       KV  +
Sbjct: 836 XKVSTYNHPDFMYISTTE--NAIKKEQVEQLVRHMNQLPIESTNKVYIIEDFEKLTVQGE 663

Query: 126 XXXXXXXEEPPAETWFFLATREPERLLATLRSRCRLHY 163
                  EEPP  T   L + +PE++L T+ SRC+  Y
Sbjct: 662 NSILKFLEEPPDNTIAILLSTKPEQILDTIHSRCQHVY 549
```

```
 gnl|CBCUMN|Pmultocida.990513.Contig705 Pasteurella multocida PM70 unfinished fragment of
          Length = 3829

 Score = 36.0 bits (81), Expect = 0.19
 Identities = 22/81 (27%), Positives = 33/81 (40%)
 Frame = +1

Query: 90   VDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXXXXXEEPPAETWFFLATREPE 149
             V+  RE+ + +        G   KV   +                     EEPP     F LAT +P+
Sbjct: 58   VEDTRELLDNVQYKPVQGRYKVYLIDEVHMLSRHSFNALLKTLEEPPEYVKFLLATTDPQ 237

Query: 150  RLLATLRSRCRLHYLAGPPEQ 170
             +L  T+ SRC    +L     +Q
Sbjct: 238  KLPITILSRCMQFHLKALEQQ 300


 gb|AB001339|SYNECHO Synechocystis PCC6803 complete genome
              Length = 3573470

 Score = 35.6 bits (80), Expect = 0.25
 Identities = 15/28 (53%), Positives = 20/28 (70%)
 Frame = -1

Query: 133      EEPPAETWFFLATREPERLLATLRSRCR 160
              EEPP     F LAT +P+R+L T+ SRC+
Sbjct: 1067285 EEPPERVVFVLATTDPQRVLPTIISRCQ 1067202


 gnl|TIGR|M.avium_5418 Mycobacterium avium unfinished fragment of complete genome
           Length = 17971

 Score = 32.5 bits (72), Expect = 2.1
 Identities = 14/28 (50%), Positives = 17/28 (60%)
 Frame = -3

Query: 29   IQALPGMGDDALIYALSRYLLCQQPQGH 56
          +   LP +GDDA+       R LL QQP GH
Sbjct: 8381 VDRLPAVGDDAVHQLARRQLLTQQPDGH 8298


 gnl|TIGR|C.crescentus_gcc_2104 Caulobacter crescentus unfinished fragment of complete ge
           Length = 826

 Score = 32.1 bits (71), Expect = 2.8
 Identities = 35/140 (25%), Positives = 54/140 (38%), Gaps = 11/140 (7%)
 Frame = -2

Query: 21  GRGHHALLIQALPGMGDDALIYALSRYLLCQQP-----------QGHKSCGHCRGCQLMQ 69
            GR  HA ++   + G+G          L+R L  +                    +G+     HCR     +
Sbjct: 822 GRIAHAFMLTGVRGVGKTTTARLLARALNYETDTVKGPSVDLTTEGY----HCRS---II 664

Query: 70  AGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVXXXXXXXXXXXXXXX 129
              G H D    L        +    VD +RE+ + +              KV   +
Sbjct: 663 EGRHMDVLEL---DAASRTKVDEMRELLDGVRYAPVEARYKVYIIDEVHMLSTAAFNALL 493

Query: 130 XXXEEPPAETWFFLATREPERLLATLRSRCR 160
         EEPP      F   AT  E   ++   T+ SRC+
Sbjct: 492 KTLEEPPPHAKFIFATTEIRKVPVTILSRCQ 400


 gnl|TIGR|V.cholerae_asm959 Vibrio cholerae unfinished fragment of complete genome
```

```
            Length = 15780

 Score = 30.5 bits (67), Expect = 8.3
 Identities = 13/41 (31%), Positives = 22/41 (52%)
 Frame = +3

Query: 219  ALAYSVPSGDWYSLLAALNHEQAPARLHWLATLLMDALKRH 259
             ++A S P+G+W   +  +    A  +  W+ATL  D L R+
Sbjct: 1191 SVALSTPNGEWGQTVKFVRRFSAQEQKEWIATLAADMLLRY 1313


CPU time:     0.64 user secs.      1.41 sys. secs         2.05 total secs.

  Database: Unfinished Actinobacillus actinomycetemcomitans
    Posted date:  Dec 30, 1998  1:59 PM
  Number of letters in database: 1,888,023
  Number of sequences in database:  537

  Database: Complete Aquifex aeolicus
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 1,551,335
  Number of sequences in database:  1

  Database: Complete Bacillus subtilis
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 4,214,814
  Number of sequences in database:  1

  Database: Unfinished Bordetella pertussis
    Posted date:  May 3, 1999  3:37 PM
  Number of letters in database: 3,987,145
  Number of sequences in database:  543

  Database: Borrelia burgdorferi
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 1,229,458
  Number of sequences in database:  12

  Database: Unfinished Campylobacter jejuni
    Posted date:  Nov 17, 1998 10:56 AM
  Number of letters in database: 1,641,480
  Number of sequences in database:  1

  Database: Complete Chlamydia trachomati
    Posted date:  Aug 14, 1998  4:20 PM
  Number of letters in database: 1,042,519
  Number of sequences in database:  1

  Database: Unfinished Chlorobium tepidum
    Posted date:  Feb 8, 1999 10:29 AM
  Number of letters in database: 2,257,254
  Number of sequences in database:  254

  Database: Unfinished Clostridium acetobutylicum
    Posted date:  Mar 31, 1999 10:56 AM
  Number of letters in database: 3,943,874
  Number of sequences in database:  1

  Database: Unfinished Caulobacter crescentus
    Posted date:  Feb 8, 1999 11:17 AM
  Number of letters in database: 4,177,031
```

```
Number of sequences in database:  3481

Database: Unfinished Chlamydia trachomatis MOPN
  Posted date:  Feb 8, 1999 11:21 AM
Number of letters in database: 1,160,971
Number of sequences in database:  624

Database: Unfinished Deinococcus radiodurans
  Posted date:  Feb 8, 1999 10:30 AM
Number of letters in database: 3,615,037
Number of sequences in database:  869

Database: Complete Escherichia coli
  Posted date:  Aug 5, 1998  9:37 AM
Number of letters in database: 4,639,221
Number of sequences in database:  1

Database: Unfinished Enterococcus faecalis
  Posted date:  Feb 8, 1999 10:30 AM
Number of letters in database: 3,209,119
Number of sequences in database:  293

Database: Complete Haemophilus influenzae Rd
  Posted date:  Aug 5, 1998  9:37 AM
Number of letters in database: 1,830,138
Number of sequences in database:  1

Database: Complete Helicobacter pylori 26695
  Posted date:  Jan 25, 1999  3:20 PM
Number of letters in database: 1,667,867
Number of sequences in database:  1

Database: Complete Helicobacter pylori J99
  Posted date:  Jan 25, 1999  3:55 PM
Number of letters in database: 1,643,831
Number of sequences in database:  1

Database: Unfinished Mycobacterium avium
  Posted date:  May 17, 1999  1:55 PM
Number of letters in database: 5,354,737
Number of sequences in database:  692

Database: Unfinished Mycobacterium bovis
  Posted date:  May 10, 1999  1:17 PM
Number of letters in database: 4,093,505
Number of sequences in database:  931

Database: Complete Mycoplasma pneumoniae
  Posted date:  Aug 5, 1998  9:37 AM
Number of letters in database: 816,394
Number of sequences in database:  1

Database: Unfinished Mycobacterium tuberculosis CSU#93
  Posted date:  Feb 8, 1999 10:30 AM
Number of letters in database: 4,306,088
Number of sequences in database:  42

Database: Complete Mycobacterium tuberculosis H37Rv
  Posted date:  Aug 14, 1998  4:20 PM
Number of letters in database: 4,411,529
Number of sequences in database:  1
```

Database: Complete Mycoplasma genitalium
Posted date: Aug 5, 1998 9:36 AM
Number of letters in database: 580,073
Number of sequences in database: 1

Database: Unfinished Neisseria gonorrhoea
Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 2,172,011
Number of sequences in database: 159

Database: Unfinished Neisseria meningitidis MC58
Posted date: Feb 8, 1999 10:30 AM
Number of letters in database: 1,406,901
Number of sequences in database: 2533

Database: Unfinished Neisseria meningitidis serogroup A
Posted date: May 3, 1999 3:38 PM
Number of letters in database: 2,166,687
Number of sequences in database: 25

Database: Unfinished Pseudomonoas aeruginosa
Posted date: Mar 15, 1999 3:11 PM
Number of letters in database: 6,246,116
Number of sequences in database: 12

Database: Unfinished Porphyromonas gingivalis W83
Posted date: May 17, 1999 1:55 PM
Number of letters in database: 2,334,787
Number of sequences in database: 12

Database: Unfinished Pasteurella multocida PM70
Posted date: May 14, 1999 2:09 PM
Number of letters in database: 4,166,549
Number of sequences in database: 3506

Database: Unfinished Pseudomonas putida
Posted date: May 10, 1999 3:21 PM
Number of letters in database: 201,388
Number of sequences in database: 391

Database: Complete Rickettsia prowazekii
Posted date: Nov 16, 1998 3:20 PM
Number of letters in database: 1,111,523
Number of sequences in database: 1

Database: Unfinished Staphylococcus aureus COL
Posted date: May 6, 1999 2:33 PM
Number of letters in database: 3,071,880
Number of sequences in database: 2177

Database: Unfinished Staphylococcus aureus
Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 733,437
Number of sequences in database: 506

Database: Unfinished Streptococcus mutans
Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 1,438,835
Number of sequences in database: 514

```
   Database: Unfinished Shewanella putrefaciens
     Posted date:  Feb 8, 1999 11:22 AM
   Number of letters in database: 5,974,789
   Number of sequences in database:  2430

   Database: Unfinished Streptococcus pyogenes
     Posted date:  Dec 30, 1998  2:00 PM
   Number of letters in database: 1,801,145
   Number of sequences in database:  181

   Database: Unfinished Streptococcus pneumoniae
     Posted date:  Feb 8, 1999 10:31 AM
   Number of letters in database: 2,114,666
   Number of sequences in database:  270

   Database: Unfinished Salmonella typhi
     Posted date:  May 3, 1999  3:38 PM
   Number of letters in database: 5,088,553
   Number of sequences in database:  185

   Database: Complete Synechocystis PCC6803
     Posted date:  Aug 5, 1998  9:36 AM
   Number of letters in database: 3,573,470
   Number of sequences in database:  1

   Database: Unfinished Thiobacillus ferrooxidans
     Posted date:  May 10, 1999  3:22 PM
   Number of letters in database: 3,488,401
   Number of sequences in database:  2870

   Database: Unfinished Thermotoga maritima
     Posted date:  Feb 8, 1999 10:31 AM
   Number of letters in database: 2,352,161
   Number of sequences in database:  948

   Database: Complete Treponema pallidum
     Posted date:  Aug 14, 1998  4:21 PM
   Number of letters in database: 1,138,011
   Number of sequences in database:  1

   Database: Unfinished Vibrio cholerae
     Posted date:  Feb 8, 1999 10:31 AM
   Number of letters in database: 4,145,671
   Number of sequences in database:  694

   Database: Unfinished Yersinia pestis
     Posted date:  May 3, 1999  3:38 PM
   Number of letters in database: 4,937,945
   Number of sequences in database:  209

Lambda     K      H
   0.322    0.137     0.00

Gapped
Lambda     K      H
   0.270   0.0470 4.94e-324


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 58759890
```

```
Number of Sequences: 537
Number of extensions: 944282
Number of successful extensions: 4967
Number of sequences better than 10.0: 144
Number of HSP's better than 10.0 without gapping: 69
Number of HSP's successfully gapped in prelim test: 7
Number of HSP's that attempted gapping in prelim test: 4565
Number of HSP's gapped (non-prelim): 857
length of query: 334
length of database: 40,975,456
effective HSP length: 48
effective length of query: 286
effective length of database: 39731536
effective search space: 11363219296
effective search space used: 11363219296
frameshift window, decay const: 50,  0.1
T: 13
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 41 (21.9 bits)
S2: 66 (30.1 bits)
```

Identities among δ (holA encoded) from different eubacteria

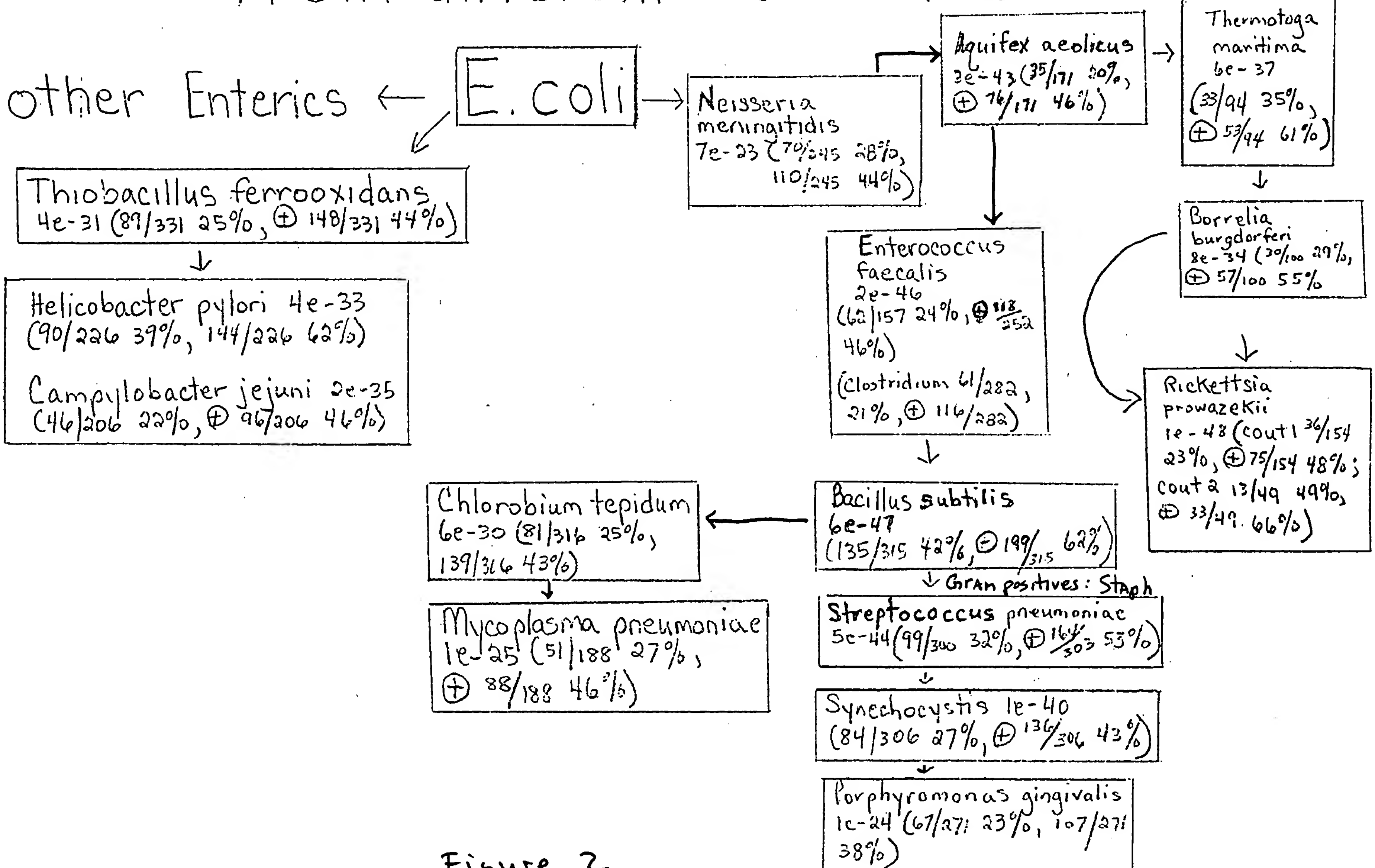


Figure 2

NCBI **BLAST Search Results** BLAST Entrez ?

WARNING: These microbial genomes from are not yet finished, and are not yet in GenBank and are not presently distributed to EMBL or DDBJ. Please see details

NOTE: This WWW-BLAST page utilizes NCBI's new gapped BLAST algorithm (Altschul et al., 1997) with the **BLASTN**, **TBLASTN**, and **TBLASTX** programs.

Commencing search, please wait for results.

You have searched a database generously provided by the Institute for Genomic Research (TIGR). Their Policy on Early Data Release is:

The Institute for Genomic Research (TIGR) releases data very rapidly to ensure that our scientific colleagues have access to information that may assist them in the search for genes and their biological function. Data releases do not constitute scientific publication, but rather provide investigators with information that may "jump-start" biological experimentation. Users of this information are encouraged to share their results with TIGR in order to improve annotation of the sequence data. Data or information may contain errors or be incomplete and should be regarded as preliminary.

TIGR asks that you acknowledge the source of information obtained from this site in any publication by including the following sentence in both the Materials and Methods and Acknowledgement sections: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" Also include the following text in the Acknowledgements, if applicable: "Sequencing of [organism name] was accomplished with support from [funding agency]." The name of the funding agency for each TIGR project can be found at http://www.tigr.org/tdb/mdb/mdb.html

Similarly, if you display this data or any information derived from it on a Web page, we ask that you prominently display the following notice on that webpage: "Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org" We request that you notify us of your electronic presentation by sending email to www@tigr.org.

TBLASTN 2.0.8 [Jan-05-1999]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query=
(343 letters)

Searching..done

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

	Score (bits)	E Value
Sequences producing significant alignments:		

```
gb|U00096|ECOLI Escherichia coli K-12 MG1655 complete genome          619  e-177
gnl|Sanger|S.typhi_Contig403 Salmonella typhi unfinished fragmen...   563  e-160
gnl|Sanger|Y.pesits_Contig765 Yersinia pestis unfinished fragmen...   447  e-125
gnl|TIGR|V.cholerae_asm937 Vibrio cholerae unfinished fragment o...   282  1e-75
gb|L42023|L42023 Haemophilus influenzae Rd complete genome            237  3e-62
gnl|OUACGT|A.actin_Contig739 Actinobacillus actinomycetemcomitan...   210  3e-58
gnl|TIGR|S.putrefaciens_gsp_230 Shewanella putrefaciens unfinish...   194  3e-49
gnl|PAGP|Paeruginosa_Contig52 Pseudomonas aeruginosa unfinished ...   139  1e-32
gnl|TIGR|t_ferrooxidans_626 Thiobacillus ferrooxidans unfinished...   126  1e-28
gnl|Sanger|B.pertussis_Contig669 Bordetella pertussis unfinished...   122  2e-27
gnl|Sanger|N.mening_Contig363 Neisseria meningitidis serogroup A...   115  2e-25
gnl|OUACGT|Ngon_Contig213 Neisseria gonorrhoeae unfinished fragm...   109  1e-23
gnl|TIGR|D.radiodurans_8857 Deinococcus radiodurans unfinished f...    38  0.064
gnl|PAGP|Paeruginosa_Contig44 Pseudomonas aeruginosa unfinished ...    31  8.2
gnl|Sanger_1765|mbovis_Contig976.0 Mycobacterium bovis unfinishe...    31  8.2
gnl|TIGR|gmt3661 Mycobacterium tuberculosis unfinished fragment ...    31  8.2
emb|AL123456|MTBH37RV Mycobacterium tuberculosis H37Rv complete ...    31  8.2

 gb|U00096|ECOLI Escherichia coli K-12 MG1655 complete genome
           Length = 4639221

 Score =  619 bits (1578), Expect = e-177
 Identities = 312/343 (90%), Positives = 312/343 (90%)
 Frame = -3

Query: 1      MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 60
              MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD
Sbjct: 670828 MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 670649

Query: 61     WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKA 120
              WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQ              IVRGNKLSKA
Sbjct: 670648 WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQLLTLTGLLHDDLLLIVRGNKLSKA 670469

Query: 121    QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA 180
              QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA
Sbjct: 670468 QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA 670289

Query: 181    LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 240
              LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG
Sbjct: 670288 LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 670109

Query: 241    SEPVIXXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
              SEPVI                  QSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ
Sbjct: 670108 SEPVILLRTLQRELLLLVNLKRQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 669929

Query: 301    AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFIDG 343
              AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFIDG
Sbjct: 669928 AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFIDG 669800


 gnl|Sanger|S.typhi_Contig403 Salmonella typhi unfinished fragment of complete genome
           Length = 36914

 Score =  563 bits (1436), Expect = e-160
 Identities = 279/343 (81%), Positives = 298/343 (86%)
 Frame = -3

Query: 1     MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 60
             MIRLYPEQLRAQLNE LRAAYLLLGNDPLLLQESQDA+R  AA+QGFEEHH F++DP+TD
Sbjct: 15489 MIRLYPEQLRAQLNEXLRAAYLLLGNDPLLLQESQDAIRLAAASQGFEEHHAFTLDPSTD 15310
```

```
Query: 61    WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKA 120
             W ++FSLCQAMSLFASRQTL+L LPENGPNAA+NEQ              IVRGNKL+KA
Sbjct: 15309 WGSLFSLCQAMSLFASRQTLVLQLPENGPNAAMNEQLATLSELLHDDLLLIVRGNKLTKA 15130

Query: 121   QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA 180
             QENAAW+TALA+RSVQV+CQTPEQAQLPRWVAARAK  NL+LDDAANQ+LCYCYEGNLLA
Sbjct: 15129 QENAAWYTALADRSVQVSCQTPEQAQLPRWVAARAKAQNLQLDDAANQLLCYCYEGNLLA 14950

Query: 181   LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 240
             LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG
Sbjct: 14949 LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 14770

Query: 241   SEPVIXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
             SEPVI                 QSAHTPLRALFDKHRVWQNRR M+G+AL RL   QLRQ
Sbjct: 14769 SEPVILLRTLQRELLLLVNLKRQSAHTPLRALFDKHRVWQNRRPMIGDALQRLHPAQLRQ 14590

Query: 301   AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFIDG 343
             AVQLLTRTE+TLKQDYGQSVWA+LEGLSLLLCHK LADVFIDG
Sbjct: 14589 AVQLLTRTEITLKQDYGQSVWADLEGLSLLLCHKALADVFIDG 14461
```

```
gnl|Sanger|Y.pesits_Contig765 Yersinia pestis unfinished fragment of complete genome
          Length = 215860

 Score =  447 bits (1138), Expect = e-125
 Identities = 223/342 (65%), Positives = 263/342 (76%)
 Frame = +3

Query: 1     MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 60
             MIR+YPEQL AQL+EGLRA YLL GN+PLLLQESQD +R+VA+   F EH +F++D +T+
Sbjct: 50067 MIRIYPEQLVAQLHEGLRACYLLCGNEPLLLQESQDHIRRVASQHDFTEHFSFALDAHTE 50246

Query: 61    WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKA 120
             W  IFSLCQA+SLFASRQTLLL  P++G  A I+EQ              I+R NKL+KA
Sbjct: 50247 WEHIFSLCQALSLFASRQTLLLSFPDSGLTAPISEQLVKLSGLLHPDILLILRANKLTKA 50426

Query: 121   QENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLA 180
             QEN+AWF AL+   V V+CQTPEQAQLPRWV+ARAK LNL +DDAA Q+LCYCYEGNLLA
Sbjct: 50427 QENSAWFKALSKNGVFVSCQTPEQAQLPRWVSARAKSLNLNVDDAAIQLLCYCYEGNLLA 50606

Query: 181   LAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEG 240
             L+QALERLSLL+PDGKLTLP+VEQAVNDAAHFTP+HW+DALLMGKSKRA HILQQL+ E
Sbjct: 50607 LSQALERLSLLYPDGKLTLPKVEQAVNDAAHFTPYHWLDALLMGKSKRAWHILQQLQQED 50786

Query: 241   SEPVIXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
             SEPVI                 Q    PLRALFD+H++WQNRR MM +AL RLS  QL+Q
Sbjct: 50787 SEPVILLRTVQRELLLLLALKRQMEQVPLRALFDQHKIWQNRRPMMTQALQRLSLQQLQQ 50966

Query: 301   AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCHKPLADVFID 342
             AV LLT+ E+ LKQDYGQS+W ELE LS+L+C K L + F D
Sbjct: 50967 AVHLLTQMEIRLKQDYGQSIWPELETLSMLMCGKTLPESFFD 51092
```

```
gnl|TIGR|V.cholerae_asm937 Vibrio cholerae unfinished fragment of complete genome
          Length = 6994

 Score =  282 bits (714), Expect = 1e-75
 Identities = 151/332 (45%), Positives = 207/332 (61%), Gaps = 1/332 (0%)
 Frame = +2

Query: 2     IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
```

```
          +R+Y E+L    L++ L    YL+ GN+PLLLQE++ A+ + A AQGF E H FS D    DW
Sbjct: 1166 MRIYAEKLAESLHKTLYPIYLVFGNEPLLLQEAKTAIEKTAQAQGFLEKHRFSADAGLDW 1345

Query: 62   NAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKAQ 121
          NA++  CQA+SLF+SRQ + + +PE+G NA    ++                +V G KL+KAQ
Sbjct: 1346 NAVYDCCQALSLFSSRQLIEIEIPESGVNAQTAKELSALVGQLHQDILLLVIGPKLTKAQ 1525

Query: 122  ENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLAL 181
          ENAAWF   LA ++   V C TPE ++LP++V   R    L L+ D   A Q+L    +EGNL AL
Sbjct: 1526 ENAAWFKTLAQQACWVNCLTPELSRLPQFVQQRCFALGLKPDAEAVQMLAQWHEGNLFAL 1705

Query: 182  AQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGS 241
          AQ+LE+L+LL+PDG LTL R+E++++    HFTP+HW+DALL GK+ RA  IL+QL LE S
Sbjct: 1706 AQSLEKLALLYPDGLLTLVRLEESLSRHNHFTPYHWMDALLEGKANRAQRILRQLMLEES 1885

Query: 242  EPVIXXXXXXXXXXXXXXXXXXQSAHT-PLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
          EP+I                         +          L +LFD++RVWQNRR +      AL RL       L +
Sbjct: 1886 EPIILIRTAQKELTQLLKWQQERQQLGNLGSLFDRYRVWQNRRPLYSAALQRLPSRALLR 2065

Query: 301  AVQLLTRTELTLKQDYGQSVWAELEGLSLLLCH 333
           V +LT+ EL  K  Y Q VW  L+ LSL  C+
Sbjct: 2066 LVGILTQAELLAKTQYEQPVWPILQQLSLECCN 2164
```

gb|L42023|L42023 Haemophilus influenzae Rd complete genome
Length = 1830138

Score = 237 bits (599), Expect = 3e-62
Identities = 133/332 (40%), Positives = 182/332 (54%), Gaps = 12/332 (3%)
Frame = +3

```
Query: 1      MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTD 60
            M R++PEQL    L +GL    YLL G DPLLL E++D + QVA   QGF+E +T   +D   TD
Sbjct: 980328 MNRIFPEQLNHHLAQGLARVYLLQGQDPLLLSETEDTICQVANLQGFDEKNTIQVDSQTD 980507

Query: 61     WNAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKA 120
            W   +    CQ++ LF S+Q L L LPEN    A + +                   I++    KL+K
Sbjct: 980508 WAQLIESCQSIGLFFSKQILSLNLPENF-TALLQKNLQELISVLHKDVLLILQVAKLAKG 980684

Query: 121    QENAAWFTALAN---RSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGN 177
             E    WF   L        ++ + CQTP       LPRWV   R K + L+ D+ A Q LCY YE N
Sbjct: 980685 IEKQTWFITLNQYEPNTILINCQTPTVENLPRWVKNRTKAMGLDADNEAIQQLCYSYENN 980864

Query: 178    LLALAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLR 237
            LLAL QAL+ L LL+PD KL     RV     V   ++ FTPF W+DALL+GK+ RA   IL+ L+
Sbjct: 980865 LLALKQALQLLDLLYPDHKLNYNRVISVVEQSSIFTPFQWIDALLVGKANRAKRILKGLQ 981044

Query: 238    LEGSEPVIXXXXXXXXXXXXXXXXXQSAH---------TPLRALFDKHRVWQNRRGMMGE 288
             E  +PVI                                  ++   FD+  ++WQNRR +
Sbjct: 981045 AEDVQPVILLRTLQRELFTLLELTKPQQRIVTTEKLPIQQIKTEFDRLKIWQNRRPLFLS 981224

Query: 289    ALNRLSQTQLRQAVQLLTRTELTLKQDYGQSVWAELEGLSLLLC 332
            A+ RL+     L + +Q L     E     KQ++     VW +L    LS+ +C
Sbjct: 981225 AIQRLTYQTLYEIIQELANIERLAKQEFSDEVWIKLADLSVKIC 981356
```

gnl|OUACGT|A.actin_Contig739 Actinobacillus actinomycetemcomitans unfinished fragment of genome
Length = 4889

Score = 210 bits (529), Expect(2) = 3e-58

```
 Identities = 124/297 (41%), Positives = 165/297 (54%), Gaps = 12/297 (4%)
 Frame = -3

Query: 33   ESQDAVRQVAAAQGFEEHHTFSIDPNTDWNAIFSLCQAMSLFASRQTLLLLLPENGPNAA 92
            ES + + Q A  +GF+E      I+ +TDWN +F   Q+M LF ++Q ++L LPEN   A
Sbjct: 2601 ESANGIYQTALQRGFDEKVELDINASTDWNDLFEPVQSMGLFFNKQLIILDLPENA-TAL 2425

Query: 93   INEQXXXXXXXXXXXXXXIVRGNKLSKAQENAAWFTALAN---RSVQVTCQTPEQAQLPR 149
            + +                I R  KL+KA E  AWF A       ++V V CQTP   QLPR
Sbjct: 2424 LQKNLSEFISLLQPDVLPIFRLAKLTKAAEKQAWFMAANQYEPQAVLVNCQTPNAEQLPR 2245

Query: 150  WVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLWPDGKLTLPRVEQAVNDA 209
            WVA RAK L L ++  A Q+LCY YE NLLAL Q L+ L LL+PD KLT  RV   V  +
Sbjct: 2244 WVANRAKMLGLSIEQEAVQLLCYSYENNLLALKQTLQLLDLLYPDRKLTFARVNSVVEQS 2065

Query: 210  AHFTPFHWVDALLMGKSKRALHILQQLRLEGSEPVIXXXXXXXXXXXXXXXXX----QSA 265
            + FTPF WVDA+L GK  RA  IL  L+ E  +P+I                     QS
Sbjct: 2064 SVFTPFQWVDAILGGKGNRARRILTGLKDEDVQPIILLRTLQRDLMTLLEISKPEQPQSL 1885

Query: 266  HTP-----LRALFDKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTRTELTLKQDYGQSV 320
             +P     LR  FD+ +VWQNRR +  +A+ RL+  +L   Q L  E  KQ++   +
Sbjct: 1884 DSPLPTDQLREQFDRLKVWQNRRSLFTQAVQRLTYRKLYLFFQQLADVERCAKQEFSDDI 1705

Query: 321  WAELEGLSL 329
            W +LE LS+
Sbjct: 1704 WQQLEDLSV 1678


 Score = 36.0 bits (81), Expect(2) = 3e-58
 Identities = 17/31 (54%), Positives = 20/31 (63%)
 Frame = -2

Query: 1    MIRLYPEQLRAQLNEGLRAAYLLLGNDPLLL 31
            M RL+PEQL + L   L   Y L+G DPLLL
Sbjct: 2698 MNRLFPEQLASSLERHLAHVYFLVGEDPLLL 2606


 gnl|TIGR|S.putrefaciens_gsp_230 Shewanella putrefaciens unfinished fragment of complete
           Length = 21837

 Score =  194 bits (489), Expect = 3e-49
 Identities = 121/341 (35%), Positives = 167/341 (48%), Gaps = 4/341 (1%)
 Frame = +2

Query: 2     IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
             +R+YP+QL    LN  L A YL+ G+DP LL+ S+D +RQ A  QGFEE       +  +W
Sbjct: 14210 MRVYPDQLSRHLNP-LHACYLIFGDDPWLLETSKDQIRQAAKRQGFEERVQLIQETGFNW 14386

Query: 62    NAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKAQ 121
              +     QAMSLF+SR+ + L LP   P A  +                I+ G KL+  Q
Sbjct: 14387 GDLTQEWQAMSLFSSRRIIELTLPSAKPGADGSAALQSLLQTPSPDVLLILEGPKLASEQ 14566

Query: 122   ENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLAL 181
              N+ WF  L +  + + C TPE  Q  RW+ +R      L L   A  +L   YEGNLLA
Sbjct: 14567 TNSKWFKTLDSLGIYLPCTTPEGDQFRRWLDSRIAHFKLNLQPDARAMLYSLYEGNLLAA 14746

Query: 182   AQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGS 241
              QA++ L LL P   +     +      D + FT F   DALL   +   A H+L QL  EG+
Sbjct: 14747 DQAMQLLQLLSPSKPIGADELSHYFEDQSRFTVFQLTDALLNNRQDSAQHMLAQLNGEGT 14926

Query: 242   E-PVIXXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQ 300
```

```
               P++                        Q+   +PL +LF KHR+W  R+ +   AL RLS  Q+
Sbjct: 14927 AMPILLWALFKELQLLLSLKSEQAQGSPLNSLFGKHRIWDKRKPLYQTALQRLSLAQIEH 15106

Query: 301    AVQLLTRTELTLKQDYGQSVWAELEGLSLLL---CHKPLADVFID 342
              +    ++ EL LKQ  G    W  L   L LL      H  LA + +D
Sbjct: 15107 MLAFASKLELNLKQ-LGHEDWTGLSHLCLLFDPKAHSHLAHINLD 15238


 gnl|PAGP|Paeruginosa_Contig52 Pseudomonas aeruginosa unfinished fragment of complete ger
              Length = 872680

 Score =  139 bits (347), Expect = 1e-32
 Identities = 106/329 (32%), Positives = 155/329 (46%), Gaps = 8/329 (2%)
 Frame = -2

Query: 2      IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
              ++L P QL    L    L    Y++ G++PLL QE+ DA+RQ      + F E    F+ + N DW
Sbjct: 245226 MKLTPAQLAKHLQGPLAPVYVVSGDEPLLCQEACDAIRQACRERDFGERQVFNAEANFDW 245047

Query: 62     NAIFSLCQAMSLFASRQTLLLLLPENGPN---AAINEQXXXXXXXXXXXXXXIVRGNKLS 118
               +       ++SLFA ++ + L LP    P      AAI ++                +     KL
Sbjct: 245046 GLLLEAGASLSLSLFAEKRLIELRLPSGKPGDKGAAILQEYLQRPPEDTVLLLGLP---KLD 244876

Query: 119    KAQENAAWFTAL--ANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEG 176
               + +      W  AL    N +   +        +  QLP+W+  R  Q  L        A +++     EG
Sbjct: 244875 GSTQKTKWAKALIDGNAAQFIQVWPVDVHQLPQWIRQRLSQAGLSASPEALELIAARVEG 244696

Query: 177    NLLALAQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQL 236
              NLLA AQ +E+L LL      ++      V+ AV D+A F   F   +DA L G++   AL IL+ L
Sbjct: 244695 NLLAAAQEIEKLKLLAEGNQIDAATVQAAVADSARFDVFGLIDAALGGEAAHALRILEGL 244516

Query: 237    RLEGSE-PVIXXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHR--VWQNRRGMMGEALNRL 293
              R EG E PVI                                  PL    F + R   VW   RR ++   AL R
Sbjct: 244515 RGEGIEPPVILWGLAREIRLLAGLSQQYGQGIPLEKAFAQARPPVWDKRRPLLTRALQRH 244336

Query: 294    SQTQLRQAVQLLTRTELTLKQDYGQSVWAELEGLSLL 330
              S ++      Q+L    +L    Q   GQ+   +    GLSLL
Sbjct: 244335 SSSRWN---QMLRDAQLIDAQIKGQAPGSPWSGLSLL 244234


 gnl|TIGR|t_ferrooxidans_626 Thiobacillus ferrooxidans unfinished fragment of complete ge
          Length = 1632

 Score =  126 bits (313), Expect = 1e-28
 Identities = 84/331 (25%), Positives = 148/331 (44%)
 Frame = -2

Query: 2    IRLYPEQLRAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDW 61
            +RL P     + L    L + Y +   ++PLLLQE++DA+    AA   GF +      +        W
Sbjct: 1184 MRLKPAHWASHLRGPLASVYGIFSDEPLLLQEAEDALMAAAAQHGFAQKQRLAQQDGGIW 1005

Query: 62   NAIFSLCQAMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKAQ 121
            +A+       A SLFA+++ LLL L              +                ++ G +    +
Sbjct: 1004 DALRDERDAGSLFAAQRVLLLRLDSPKVPKEASAALQYWLASPPPDALLVLSGPRPDASI 825

Query: 122  ENAAWFTALANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLAL 181
            +  AWF  +     +    PE      PRWV   R +    ++ D AA Q+L        GNL A
Sbjct: 824  QKTAWFKGIETHGHTLLLYRPEGQDWPRWVEQRLRAAGMQADSAAVQLLTDLSAGNLGAC 645

Query: 182  AQALERLSLLWPDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGS 241
              QA++RL  ++P  ++     +    + D++ FT +    DA+L G+++   LH+L +LR
```

```
Sbjct: 644   HQAIQRLQQVYPGQRIDAVAIRAVLADSSQFTIYDLADAVLRGETEHMLHMLDRLRNGDG 465

Query: 242   EPVIXXXXXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQA 301
             EP +                   ++      + A F ++R++   R+G +   A   RL+++ L+
Sbjct: 464   EPAL--CLWVLHKDLRLLAELRAGGVDVDAFFRQNRIFPPRQGWLRTAARRLTRSGLQXG 291

Query: 302   VQLLTRTELTLKQDYGQSVWAELEGLSLLLC 332
             ++       +  +K         VW  L  L L +C
Sbjct: 290   IKDCLAIDARIKGQDPTPVWPALTDLCLRMC 198
```

gnl|Sanger|B.pertussis_Contig669 Bordetella pertussis unfinished fragment of complete ge
Length = 24999

Score = 122 bits (302), Expect = 2e-27
Identities = 89/313 (28%), Positives = 142/313 (44%), Gaps = 12/313 (3%)
Frame = -1

```
Query: 17    LRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDWNAIFSLCQAMSLFAS 76
             L     Y + G++PLL+ E+ DA+R   A A G+ +    +    +D    +DW+A+ +   Q++SLF
Sbjct: 3297  LAPLYTVSGDEPLLVTEAADAIRAAARAAGYTDRTSMVMDARSDWSAVAAATQSVSLFGD 3118

Query: 77    RQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXX----IVRGNKLSKAQENAAWFTALAN 132
             R+ L L +P    P  +  E                          +V    +L KA    + W    LA
Sbjct: 3117  RRLLELKIPTGKPGKSGGEMLARLADQARDQADADTLVVVALPRLDKATRESKWAQXLAR 2938

Query: 133   RSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLW 192
               V          E+ +LP W+   R   +        D A   Q +     EGN LA   Q +++L LL+
Sbjct: 2937  GGVMADIANVERGRLPAWIGMRLGRXXQRADTATLQWMADKVEGNXLAAHQEIQKLGLLY 2758

Query: 193   PDGKLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGSE-PVIXXXXXX 251
             P+G+L       VE+AV     A  +    F     DA+L G + R + +L    LR EG      P++
Sbjct: 2757  PEGQLXAEDVERAVLXVARYDVFGLRDAMLAGDTARTVRMLXGLRAEGEALPLVLWAVGE 2578

Query: 252   XXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTRTELT 311
                               +            AL    + R++          +   +AL  R++              AVQ          +
Sbjct: 2577  EIRLLARVAQARQQGQDAGALMRRLRIFGAHERLALQALGRVAPGAWPAAVQHAHEVDRL 2398

Query: 312   LKQDYGQSV-------WAELEGLSL 329
             +K     G SV              W E+    L+L
Sbjct: 2397  IK---GLSVPGRLADPWEEMTRLAL 2332
```

gnl|Sanger|N.mening_Contig363 Neisseria meningitidis serogroup A unfinished fragment of
genome
Length = 76426

Score = 115 bits (286), Expect = 2e-25
Identities = 81/322 (25%), Positives = 137/322 (42%), Gaps = 2/322 (0%)
Frame = +2

```
Query: 10    RAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDWNAIFSLCQ 69
             R     +    L+     Y++   G  + LL      E+ DA+R     A    QG+           ++ D        DWN +
Sbjct: 62483 RIDTDAPLKPLYVIHGEEELLRIEALDALRAAAKKQGYLNREVYTADNAFDWNELLQTAG 62662

Query: 70    AMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKAQENAAWFTA 129
             +    LFA   + L L +P     P         E                        +V      KL  K Q     + WF  A
Sbjct: 62663 SAGLFADLKLLELHIPNGKPGKTGGEALQDFAARLPEDTVTLVLLPKLEKTQLQSKWFAA 62842

Query: 130   LANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLS 189
             LA  +            +        A  LP+W++    R    ++  L  ++    A    +           EGNLLA    Q  +++L
```

```
Sbjct: 62843 LAAKGEVWEAKPVGAAALPQWIRGRLDKIGLGIEADALALFAERVEGNLLAARQEIDKLG 63022

Query: 190   LLWPDG-KLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGSEPVIXX- 247
             LL+P G  + +   + AV + A F  F    A + G   R   +L  LR EG EPV+
Sbjct: 63023 LLYPKGHTVNIDEAQTAVANVARFDAFQLAGAWMKGDVLRVCRLLDGLREEGEEPVLLLW 63202

Query: 248   XXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTR 307
                                  ++++ +  R+W +++ +   A+ R+S  +L  A++   +
Sbjct: 63203 AVAEDVRTLIRLAAALKQGQSIQSVRNSLRLWGDKQTLAPLAVKRISVVRLLDALKTCAQ 63382

Query: 308   TELTLKQDYGQSVWAELEGLSLLL 331
              +  +K       W   + L + L
Sbjct: 63383 IDRIIKGAEEGDAWTVFKRLVVSL 63454


 gnl|OUACGT|Ngon_Contig213 Neisseria gonorrhoeae unfinished fragment of complete genome
             Length = 41162

 Score =  109 bits (271), Expect = 1e-23
 Identities = 78/322 (24%), Positives = 136/322 (42%), Gaps = 2/322 (0%)
 Frame = -3

Query: 10    RAQLNEGLRAAYLLLGNDPLLLQESQDAVRQVAAAQGFEEHHTFSIDPNTDWNAIFSLCQ 69
             R   +  L+  Y++ G + LL  E+ DA+R  A  QG+     ++ D + DWN +
Sbjct: 12126 RIDTDAPLKPLYVIHGEEELLRIEAVDALRAAAKKQGYLNREAYTADASFDWNELLQTAG 11947

Query: 70    AMSLFASRQTLLLLLPENGPNAAINEQXXXXXXXXXXXXXXIVRGNKLSKAQENAAWFTA 129
                LFA  + L L +P   P     E               +V   KL K +  + WF A
Sbjct: 11946 NAGLFADLKLLELHIPNGKPGKNGGEALQDFAARLPEDTVTLVLLPKLEKTRLQSKWFAA 11767

Query: 130   LANRSVQVTCQTPEQAQLPRWVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLS 189
             LA +      +    A LP+W+  R  ++ L ++  A  +     EGNLLA  Q +++L+
Sbjct: 11766 LAAKGEVWEAKPVGAAALPQWIRGRLDKIGLGIEADALALFAERVEGNLLAARQEIDKLA 11587

Query: 190   LLWPDG-KLTLPRVEQAVNDAAHFTPFHWVDALLMGKSKRALHILQQLRLEGSEPVIXX- 247
             LL+P G  + +   + AV + A F  F    A +     R   +L  L  EG EPV+
Sbjct: 11586 LLYPKGHAVYIDEAQTAVANVARFDAFQLAGAWMKADVPRVCRLLDGLEEEGEEPVLLLW 11407

Query: 248   XXXXXXXXXXXXXXXQSAHTPLRALFDKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTR 307
                                  ++++ +  R+W +++ +   A+ R+S  +L  A++   +
Sbjct: 11406 AVAEDVRTLIRLAAALKQGQSIQSVRNSLRLWGDKQTLAPLAVKRISVVRLLDALKTCAQ 11227

Query: 308   TELTLKQDYGQSVWAELEGLSLLL 331
              +  +K       W   + L + L
Sbjct: 11226 IDRIIKGAEDGDAWTVFKQLVVSL 11155


 gnl|TIGR|D.radiodurans_8857 Deinococcus radiodurans unfinished fragment of complete genc
             Length = 22105

 Score = 37.5 bits (85), Expect = 0.064
 Identities = 28/94 (29%), Positives = 42/94 (43%)
 Frame = -1

Query: 150   WVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLWPDGKLTLPRVEQAVNDA 209
             WV  RAK++ L L+  A   L   +  +L  +A  L +L LL   G L   RV+  V
Sbjct: 14218 WVVTRAKKMGLRLERDAASYLAEVFGADLAGIAGELNKLELL--GGALNRERVQGIVGRD 14045

Query: 210   AHFTPFHWVDALLMGKSKRALHILQQLRLEGSEP 243
                  F  + A   G+   A+  L++L   G +P
Sbjct: 14044 PPGDSFAMLGAATAGRPGEAVLQLRRLLGSGEDP 13943
```

```
 gnl|PAGP|Paeruginosa_Contig44 Pseudomonas aeruginosa unfinished fragment of complete ger
             Length = 203793

 Score = 30.5 bits (67), Expect = 8.2
 Identities = 19/54 (35%), Positives = 25/54 (46%)
 Frame = +3

Query: 274    DKHRVWQNRRGMMGEALNRLSQTQLRQAVQLLTRTELTLKQDYGQSVWAELEGL 327
               D    + Q R    +G   L +L QTQ      V LL    ++        Y     V+A LEGL
Sbjct: 157899 DGEAIAQLRTDELGGLLRKLRQTQQMALVGLLRNQDVATSLGYLARVYARLEGL 158060

 gnl|Sanger_1765|mbovis_Contig976.0 Mycobacterium bovis unfinished fragment of complete g
           Length = 7357

 Score = 30.5 bits (67), Expect = 8.2
 Identities = 20/62 (32%), Positives = 31/62 (49%)
 Frame = -2

Query: 150  WVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLWPDGKLTLPRVEQAVNDA 209
             W A     QL + +D AA QV       E   +L ++   LE++ ++   D + T PRV+Q
Sbjct: 3981 WGANAGSQLQVFVD-AAGQVPQPVIENRVLLVSDPLEQIPVV*DDDQRTRPRVKQVFGRR 3805

Query: 210  AH 211
              H
Sbjct: 3804 QH 3799

 gnl|TIGR|gmt3661 Mycobacterium tuberculosis unfinished fragment of complete genome
             Length = 132053

 Score = 30.5 bits (67), Expect = 8.2
 Identities = 20/62 (32%), Positives = 31/62 (49%)
 Frame = +2

Query: 150   WVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLWPDGKLTLPRVEQAVNDA 209
              W A     QL + +D AA QV       E   +L ++   LE++ ++   D + T PRV+Q
Sbjct: 28898 WGANAGSQLQVFVD-AAGQVPQPVIENRVLLVSDPLEQIPVV*DDDQRTRPRVKQVFGRR 29074

Query: 210   AH 211
               H
Sbjct: 29075 QH 29080

 emb|AL123456|MTBH37RV Mycobacterium tuberculosis H37Rv complete genome
               Length = 4411529

 Score = 30.5 bits (67), Expect = 8.2
 Identities = 20/62 (32%), Positives = 31/62 (49%)
 Frame = +2

Query: 150     WVAARAKQLNLELDDAANQVLCYCYEGNLLALAQALERLSLLWPDGKLTLPRVEQAVNDA 209
                W A     QL + +D AA QV       E   +L ++   LE++ ++   D + T PRV+Q
Sbjct: 1893830 WGANAGSQLQVFVD-AAGQVPQPVIENRVLLVSDPLEQIPVV*DDDQRTRPRVKQVFGRR 1894006

Query: 210     AH 211
                 H
Sbjct: 1894007 QH 1894012
```

```
CPU time:     0.27 user secs.      0.60 sys. secs        0.87 total secs.

  Database: Unfinished Actinobacillus actinomycetemcomitans
    Posted date:  Dec 30, 1998  1:59 PM
  Number of letters in database: 1,888,023
  Number of sequences in database:  537

  Database: Complete Aquifex aeolicus
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 1,551,335
  Number of sequences in database:  1

  Database: Complete Bacillus subtilis
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 4,214,814
  Number of sequences in database:  1

  Database: Unfinished Bordetella pertussis
    Posted date:  May 3, 1999  3:37 PM
  Number of letters in database: 3,987,145
  Number of sequences in database:  543

  Database: Borrelia burgdorferi
    Posted date:  Aug 5, 1998  9:38 AM
  Number of letters in database: 1,229,458
  Number of sequences in database:  12

  Database: Unfinished Campylobacter jejuni
    Posted date:  Nov 17, 1998 10:56 AM
  Number of letters in database: 1,641,480
  Number of sequences in database:  1

  Database: Complete Chlamydia trachomati
    Posted date:  Aug 14, 1998  4:20 PM
  Number of letters in database: 1,042,519
  Number of sequences in database:  1

  Database: Unfinished Chlorobium tepidum
    Posted date:  Feb 8, 1999 10:29 AM
  Number of letters in database: 2,257,254
  Number of sequences in database:  254

  Database: Unfinished Clostridium acetobutylicum
    Posted date:  Mar 31, 1999 10:56 AM
  Number of letters in database: 3,943,874
  Number of sequences in database:  1

  Database: Unfinished Caulobacter crescentus
    Posted date:  Feb 8, 1999 11:17 AM
  Number of letters in database: 4,177,031
  Number of sequences in database:  3481

  Database: Unfinished Chlamydia trachomatis MOPN
    Posted date:  Feb 8, 1999 11:21 AM
  Number of letters in database: 1,160,971
  Number of sequences in database:  624

  Database: Unfinished Deinococcus radiodurans
    Posted date:  Feb 8, 1999 10:30 AM
  Number of letters in database: 3,615,037
```

Number of sequences in database: 869

Database: Complete Escherichia coli
  Posted date: Aug 5, 1998 9:37 AM
Number of letters in database: 4,639,221
Number of sequences in database: 1

Database: Unfinished Enterococcus faecalis
  Posted date: Feb 8, 1999 10:30 AM
Number of letters in database: 3,209,119
Number of sequences in database: 293

Database: Complete Haemophilus influenzae Rd
  Posted date: Aug 5, 1998 9:37 AM
Number of letters in database: 1,830,138
Number of sequences in database: 1

Database: Complete Helicobacter pylori 26695
  Posted date: Jan 25, 1999 3:20 PM
Number of letters in database: 1,667,867
Number of sequences in database: 1

Database: Complete Helicobacter pylori J99
  Posted date: Jan 25, 1999 3:55 PM
Number of letters in database: 1,643,831
Number of sequences in database: 1

Database: Unfinished Mycobacterium avium
  Posted date: May 17, 1999 1:55 PM
Number of letters in database: 5,354,737
Number of sequences in database: 692

Database: Unfinished Mycobacterium bovis
  Posted date: May 10, 1999 1:17 PM
Number of letters in database: 4,093,505
Number of sequences in database: 931

Database: Complete Mycoplasma pneumoniae
  Posted date: Aug 5, 1998 9:37 AM
Number of letters in database: 816,394
Number of sequences in database: 1

Database: Unfinished Mycobacterium tuberculosis CSU#93
  Posted date: Feb 8, 1999 10:30 AM
Number of letters in database: 4,306,088
Number of sequences in database: 42

Database: Complete Mycobacterium tuberculosis H37Rv
  Posted date: Aug 14, 1998 4:20 PM
Number of letters in database: 4,411,529
Number of sequences in database: 1

Database: Complete Mycoplasma genitalium
  Posted date: Aug 5, 1998 9:36 AM
Number of letters in database: 580,073
Number of sequences in database: 1

Database: Unfinished Neisseria gonorrhoea
  Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 2,172,011
Number of sequences in database: 159

Database: Unfinished Neisseria meningitidis MC58
  Posted date: Feb 8, 1999 10:30 AM
Number of letters in database: 1,406,901
Number of sequences in database: 2533

Database: Unfinished Neisseria meningitidis serogroup A
  Posted date: May 3, 1999 3:38 PM
Number of letters in database: 2,166,687
Number of sequences in database: 25

Database: Unfinished Pseudomonoas aeruginosa
  Posted date: Mar 15, 1999 3:11 PM
Number of letters in database: 6,246,116
Number of sequences in database: 12

Database: Unfinished Porphyromonas gingivalis W83
  Posted date: May 17, 1999 1:55 PM
Number of letters in database: 2,334,787
Number of sequences in database: 12

Database: Unfinished Pasteurella multocida PM70
  Posted date: Jun 4, 1999 9:26 AM
Number of letters in database: 2,034,447
Number of sequences in database: 644

Database: Unfinished Pseudomonas putida
  Posted date: May 10, 1999 3:21 PM
Number of letters in database: 201,388
Number of sequences in database: 391

Database: Complete Rickettsia prowazekii
  Posted date: Nov 16, 1998 3:20 PM
Number of letters in database: 1,111,523
Number of sequences in database: 1

Database: Unfinished Staphylococcus aureus COL
  Posted date: May 6, 1999 2:33 PM
Number of letters in database: 3,071,880
Number of sequences in database: 2177

Database: Unfinished Staphylococcus aureus
  Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 733,437
Number of sequences in database: 506

Database: Unfinished Streptococcus mutans
  Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 1,438,835
Number of sequences in database: 514

Database: Unfinished Shewanella putrefaciens
  Posted date: Feb 8, 1999 11:22 AM
Number of letters in database: 5,974,789
Number of sequences in database: 2430

Database: Unfinished Streptococcus pyogenes
  Posted date: Dec 30, 1998 2:00 PM
Number of letters in database: 1,801,145
Number of sequences in database: 181

```
   Database: Unfinished Streptococcus pneumoniae
     Posted date:  Feb 8, 1999 10:31 AM
   Number of letters in database: 2,114,666
   Number of sequences in database:  270

   Database: Unfinished Salmonella typhi
     Posted date:  May 3, 1999  3:38 PM
   Number of letters in database: 5,088,553
   Number of sequences in database:  185

   Database: Complete Synechocystis PCC6803
     Posted date:  Aug 5, 1998  9:36 AM
   Number of letters in database: 3,573,470
   Number of sequences in database:  1

   Database: Unfinished Thiobacillus ferrooxidans
     Posted date:  May 10, 1999  3:22 PM
   Number of letters in database: 3,488,401
   Number of sequences in database:  2870

   Database: Unfinished Thermotoga maritima
     Posted date:  Feb 8, 1999 10:31 AM
   Number of letters in database: 2,352,161
   Number of sequences in database:  948

   Database: Complete Treponema pallidum
     Posted date:  Aug 14, 1998  4:21 PM
   Number of letters in database: 1,138,011
   Number of sequences in database:  1

   Database: Unfinished Vibrio cholerae
     Posted date:  Feb 8, 1999 10:31 AM
   Number of letters in database: 4,145,671
   Number of sequences in database:  694

   Database: Unfinished Yersinia pestis
     Posted date:  May 3, 1999  3:38 PM
   Number of letters in database: 4,937,945
   Number of sequences in database:  209

Lambda     K      H
   0.321    0.134     0.00

Gapped
Lambda     K      H
   0.270   0.0470 4.94e-324


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 44868843
Number of Sequences: 537
Number of extensions: 569311
Number of successful extensions: 2852
Number of sequences better than 10.0: 34
Number of HSP's better than 10.0 without gapping: 15
Number of HSP's successfully gapped in prelim test: 6
Number of HSP's that attempted gapping in prelim test: 2763
Number of HSP's gapped (non-prelim): 198
length of query: 343
length of database: 40,264,755
```

```
effective HSP length: 53
effective length of query: 290
effective length of database: 39042946
effective search space: 11322454340
effective search space used: 11322454340
frameshift window, decay const: 50,  0.1
T: 13
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 41 (21.9 bits)
S2: 66 (30.1 bits)
```